HP ProCurve project

CERN openlab II quarterly review

9 October 2007

Ryszard Jurga

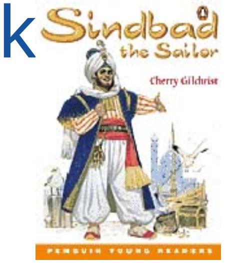
Milosz Hulboj

Outline

- Project definition
- Update on the project
- Short term goals
- Medium term goals

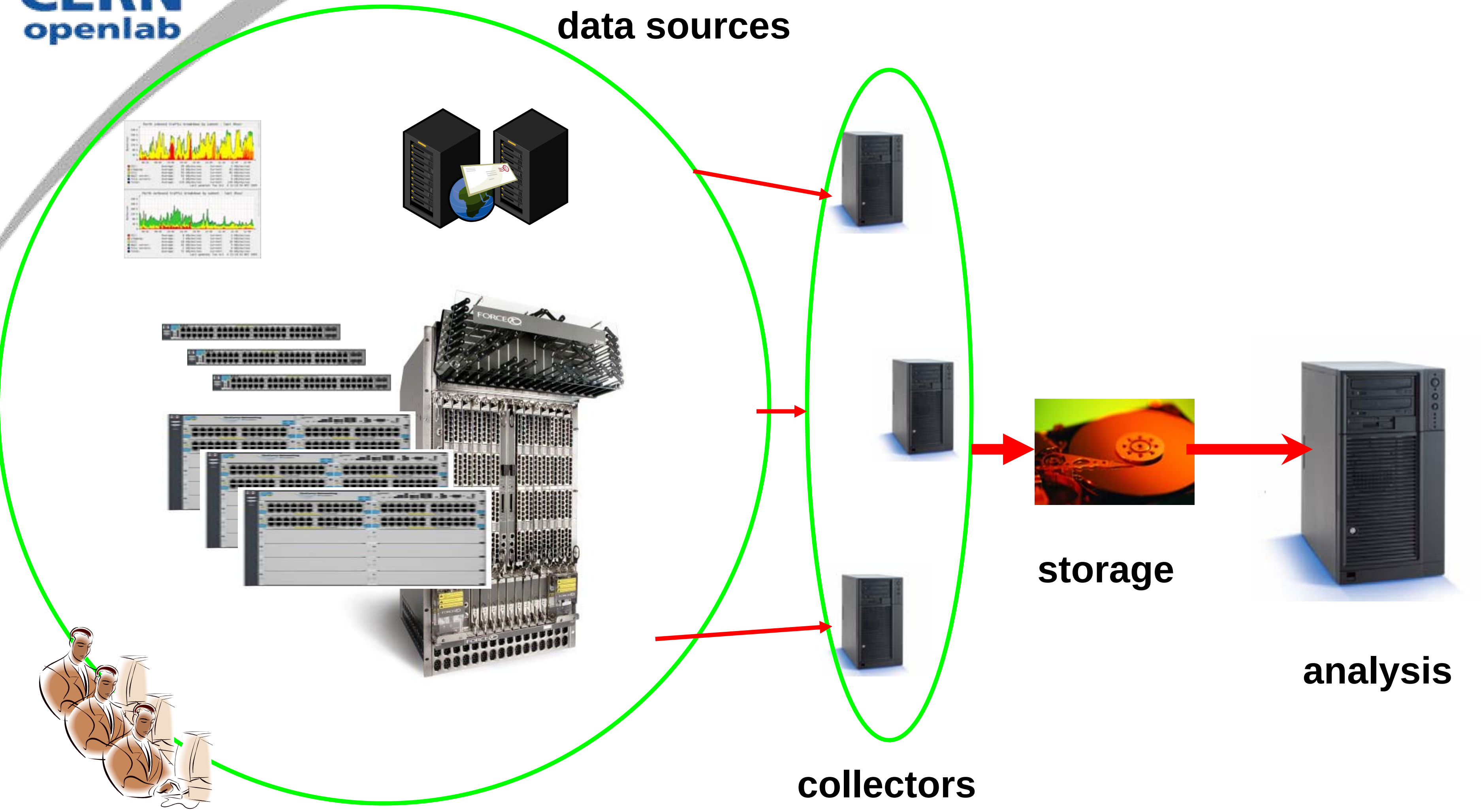
Definition

- **CINBAD: C**ern **I**nvestigation of **N**etwork **B**ehavior **A**nomaly **D**etection



- The project goal is to understand the behaviour of large computer networks (10’000+ nodes) in High Performance Computing or large Campus installations to be able to:
  - Detect traffic anomalies in the system
  - Be able to perform trend analysis
  - Automatically take counter measures
  - Provide post-mortem analysis facilities

Basic Principle



data sources

collectors

storage

analysis

Network anomaly

- Definition of „anomaly”
  - a deviation from the normal traffic pattern
  - something that differs from the expectation
  - other definitions

- Network anomaly
  - Natural
    - Misconfigured devices, system overloaded, bad cabling, etc...
  - Intentional
    - Malicious – caused by attacker or virus/worm

Data sources

- Network data sources
  - sFlow, Netflow, SNMP, RMON, probes, etc.
- Configuration data, topology
- Servers logs
  - DNS, DHCP, etc.
- Monitoring systems
  - alerts
- Human reports
  - network operator reports, user complains
- others

Data collection

- How much data do we need?
  - How much details about network do we need?
    - a port level, device, sub-net, network
  - Initially collect as much as we can
  - Later on, determine the minimal amount of data

- Distributed architecture
  - aggregate reports from collectors
  - store data in the database

- What data should be stored?
  - How to store data from different sources?

Data analysis

- Determine a baseline
  - the network patterns on different time of day, hour, week, month etc...
  - configuration changes, new applications
- Time synchronization
- Detect an anomaly
  - as a distance from the baseline
  - identify a potential source
  - fixing
  - accuracy

Update on the project

- Started on 1st July
- Training at HP Roseville, CA
  - sFlow
    - an industry standard (RFC-3176)
    - derived from **the collaboration between HP**, the University of Geneva **and CERN in 1991**
    - is based on randomly sampling one out of every N packets
    - packet header with some additional data
    - distributed agents and collectors
    - widely supported by HP ProCurve network equipment
  - PCM
    - ProCurve network management tool
  - Network anomalies (what/how/where)
  - Network hardware internals

Initial ideas

- Two brainstorm sessions
  - initially collect and store as much data as we can
    - from different network layers (by sFlow, NetFlow, SNMP, RMON, …)
    - from applications (i.e., logs from DNS, DHCP,…)
    - from network probes
    - other sources:
      - configurations and topology changes, network operators and user reports
  - more dedicated sessions are scheduled

Initial ideas cd

- Survey the network management techniques in use, in particular at CERN and in HP ProCurve
  - PCM, CERN network infrastructure, LANDB

- Web based survey of anomaly detection techniques
  - packets and flows sampling

Short term goals

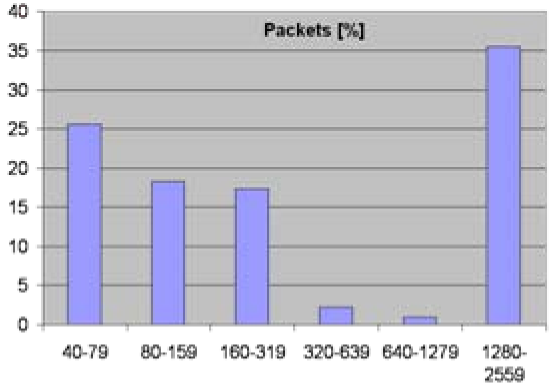
- Examine the sFlow sampling behavior
  - protocol investigation
  - device configuration,
  - sFlow limits,
  - network traffic generators,
  - etc...

- Check the behaviour of various packet capture techniques
  - Berkley sockets (udp, raw),
  - libpcap,
  - boost.asio,
  - PF_RING socket

- Portable threading library investigations
  - Boost.Thread
  - Intel TBB

Short term goals

- Set up initial traffic collection on a network device
  - real data from a production device
  - determine some real statistics in order to define initial sampling rates which do not affect the performance of devices
  - find out what could be interesting for the future investigations .e., the most popular protocols



Packet size distribution from 4 days long sampling at the rate 1/8192

Medium-term goals

- Identify and understand the sources of information available in the network infrastructure
- Perform an analysis of large-scale network data collection
- Initial implementation of a prototype of the data collector
- Investigate and propose a scalable data collector architecture
- Define structures for efficient storage and retrieval of large-scale network data
- Begin collecting network data for analysis