CERN openlab
Major Review Meeting

18th September 2008

Milosz Marian Hulboj - CERN/Procurve

Ryszard Erazm Jurga - CERN/Procurve

Agenda

- Network Anomalies
  - Definition and detection methods
- Scalable sFlow Collector
  - Implementation
  - Wide scale data collection
- First achievements
- Conclusions and Future Plans

Anomaly Definition (1)

- Anomalies are a fact in computer networks
- Anomaly definition is very domain specific:

Network faults	Malicious attacks	Viruses/worms
Misconfiguration	…	…

- But there is a common denominator:
  - *“Anomaly is a deviation of the system from the normal (expected) behaviour (baseline)”*
  - *“Normal behaviour (baseline) is not stationary and is not always easy to define”*
  - *“Anomalies are not necessarily easy to detect”*

Anomaly Definition (2)

- Just a few examples of anomalies:
  - Unauthorised DHCP server (either malicious or accidental)
  - NAT (not allowed at CERN)
  - Spreading worms/viruses
  - Exploits (attacker trying to exploit vulnerabilities)

- Examples of potential anomaly indicators:
  - TCP SYN packets without corresponding ACK
  - IP fan-out and fan-in (what about servers – i.e. DNS?)
  - Unwanted protocols on a given subnet (packets ‘*that should not be there*’)

Signature Based Detection Methods

- Perform well against known problems
- Can provide detailed information about the anomaly
- Tend to have low false positive rate

- Do not work against unknown anomalies
- Require up-to-date database of known signatures
- Numerous practical applications: antivirus software, IDS software

- Example: Signature found at W32.Netsky.p binary

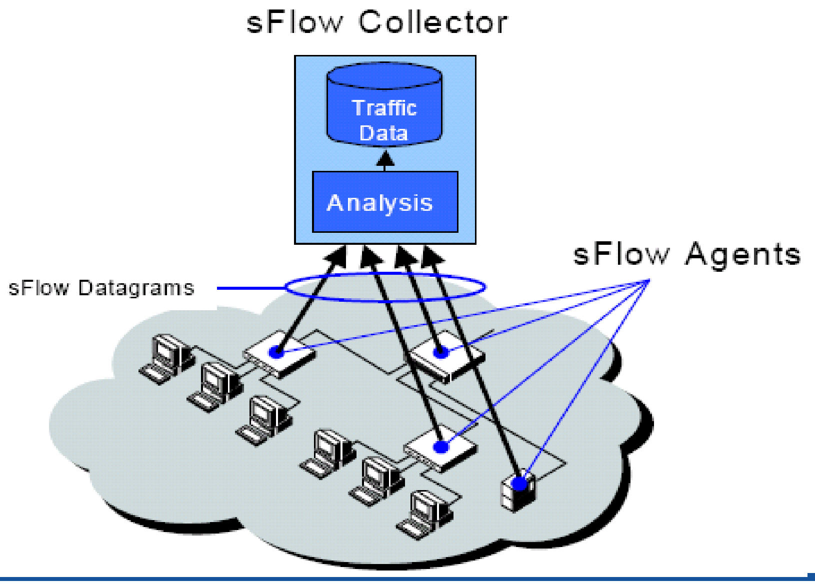
```
00000760  E7 6F 8C 88 3A 79 B3 9D 9D 52 44 AD 62 61 3D 8F   ço██:y³██RD-ba=█
00000770  98 6D 4C 07 C2 00 E5 4C 48 F0 91 4E EB 87 89 77   █mL█Â.åLHð´Në██w
00000780  7E E0 83 B1 94 94 CC E9 F5 97 97 53 95 5C 95 AF   ~à█±██Ìéõ██S█\█¯
00000790  C6 40 C5 CA AC 25 8E 47 F1 5D 0E 9F BB CB A6 67   Æ@ÅÊ¬%█Gñ]██»Ë¦g
000007A0  DB 44 E8 D2 48 3B 8F 76 CB 9E E1 53 FB FB 41 11   ÛDèÒH;█vË█áSûûA█
```

Statistical Detection Methods

- Learn the "normal behaviour" from network measurements
- Can continuously update the "normal baseline"
- Can detect new, unknown anomalies
- Selection of suitable input variables is needed
  - Many anomalies are within "normal" bounds for most of the metrics
- May be subject to attack
  - Attempt to force false negatives to occur – i.e. "boil the frog"
- Detection Rate vs False Positive Ration tradeoff
  - False positives are very costly
- Poor anomaly type identification
  - Is it a flash crowd or DDoS attack?
  - Very important issue for the real life usage

Main Datasource - sFlow

- Multi-vendor standard for passive network monitoring
- Complete packet header information
- Some SNMP counters information
- Raw sFlow data is not suitable for most types of analysis
- Conversion to a form suitable for analysis is needed

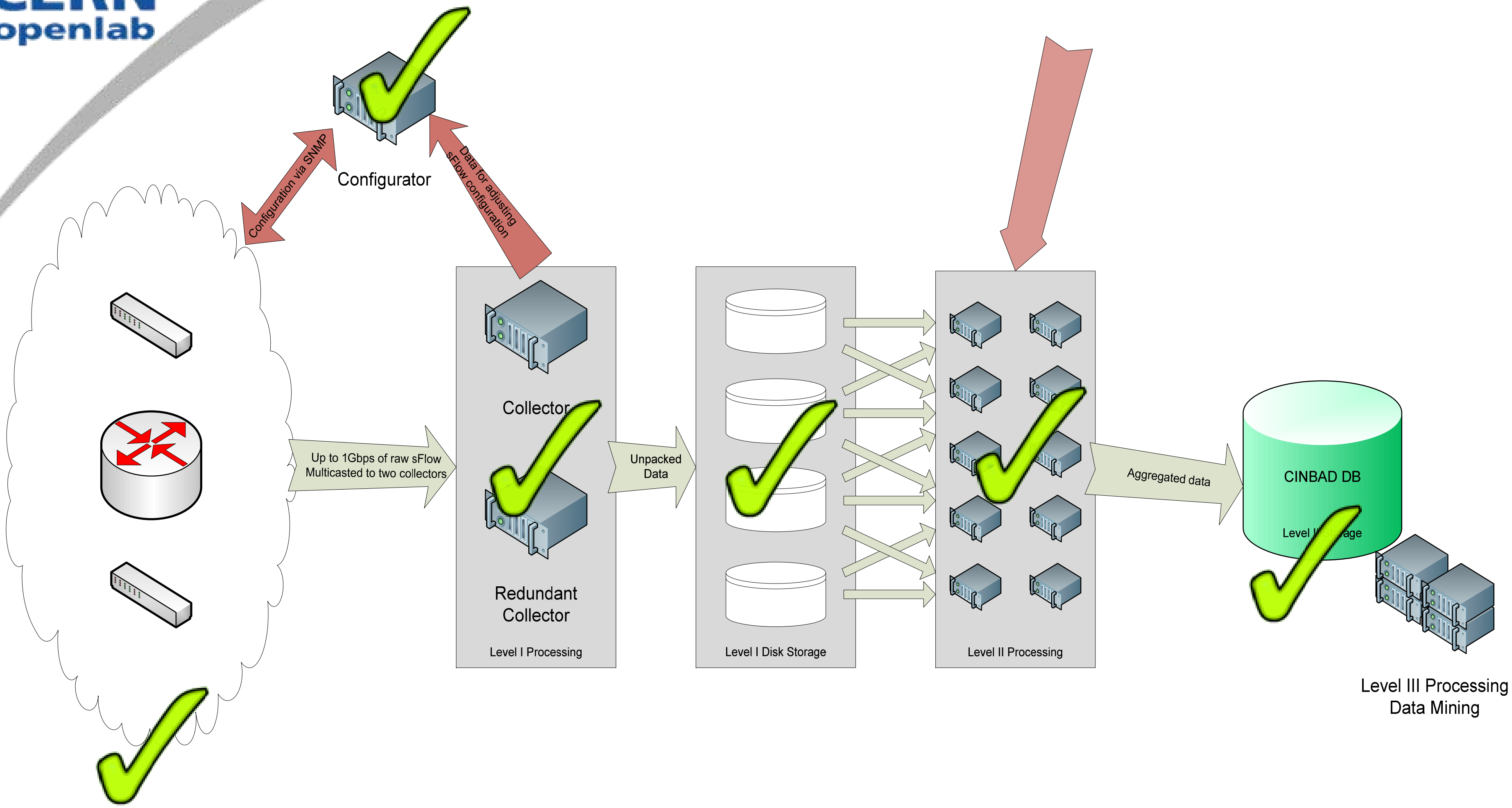


sFlow Data Collector Design

- Huge amount of raw sFlow data
  - Estimated amount: 300’000 samples/second

- Survey on data acquisition @ CERN:
  - Oracle users: Lemon, PVSS
  - LHC experiments experts consulted:
    - High performance data storage
    - Data format and representation
    - Analysis principles
- Conclusion: follow a two level strategy

Collector Implementation



Layer I Storage

- sFlow datagram tree-like format is not ideal
- Our main wishes:
  - Fast direct access to all sample elements
  - Have all the needed data in one place
  - Avoid multiple parsing of the sFlow tree
- Thus we have decided to flatten the sFlow information:
  - Raw headers stored in pcap compatible format
  - Metadata stored in separate file
  - Minimal space overhead introduced

Special tools developed for filtering the data.
Already found some interesting results!

Nataly Basha (openlab summer student) contribution

Layer II Storage

- Oracle as a long-term storage

- What should be stored:
  - We want to store the data for a long time
  - We want to store as much useful information as possible

- Currently we are storing some basic data aggregates (flow information):
  - At the L2 level (Ethernet, LLC)
  - At the L3 level (IP)
  - At the L4 level for certain protocols (TCP, UDP, ICMP)

Wide Scale Data Collection

- CINBAD sFlow data collector worked well
  - run on the Intel Dual Quad-Core server with 16GB RAM and 2TB storage
- Data collected:
  - Day1: 186 devices, over 20GB data
  - Day2: 438 devices, over 70GB data
  - Additionally received sFlow data from ATLAS experiment (over 1.5TB)
- The system has been running as expected
- Minor issues with old firmware versions

CERN openlab

First achievements

- Detected "anomalies"
  - strange device (Ethernet-to-serial hub sending any broadcasts)
  - external DNS users and strange traffic on port 53
  - the security team activities in the network

And all that just within this "small" amount of data we have from the two days testing!

- Triggered actions
  - security team decided to block the traffic to outside DNS servers
  - A policy regarding TOR and proxies usage at CERN will be prepared

Conclusions

- We have implemented working system for on-line collection and processing of the sFlow data

- We obtained encouraging initial results of data analysis

- We continue to collaborate with many parties at CERN:
  - IT-CS group
  - CERN security team
  - ATLAS Network Team

Future Plans

- Investigate anomalies related to DHCP and NAT
- How much did we miss because of the our data aggregation?
- Automate the detection process for identified types of anomalies
- Look for more anomalies
  - extend our current set of data aggregates
  - try to use machine learning methods – automate the process