CINBAD

CERN/HP ProCurve Joint Project on Networking

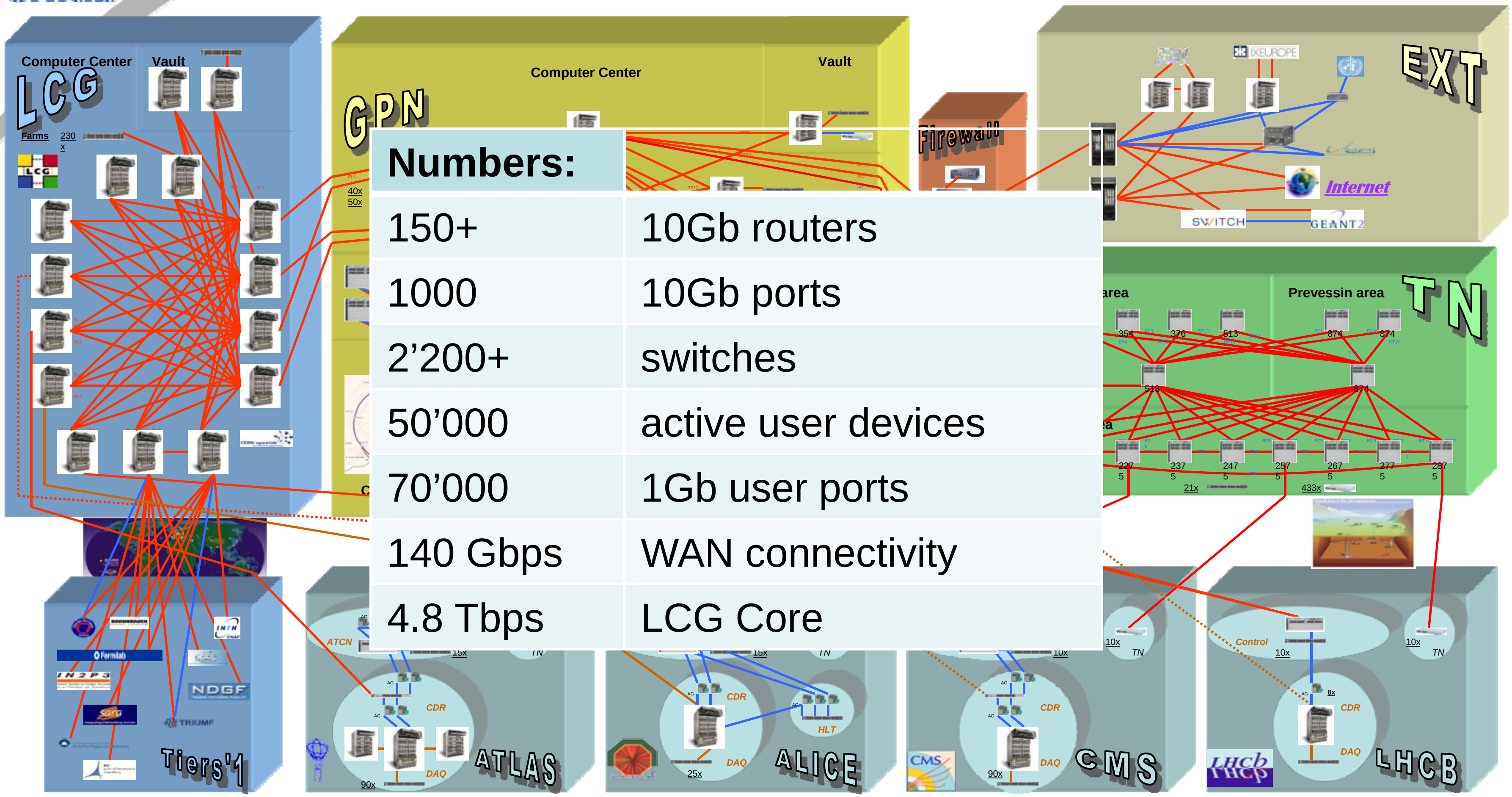
Post-C5 meeting,
12 June 2009
(hepix, 26 May 2009)

Ryszard Erazm Jurga - CERN
Milosz Marian Hulboj - CERN

Outline

- Introduction to CERN network
- CINBAD project and its goals
- Data - sources, collection and analysis
- Results and conclusions

Simplified overall CERN campus network topology



Numbers:	
150+	10Gb routers
1000	10Gb ports
2'200+	switches
50'000	active user devices
70'000	1Gb user ports
140 Gbps	WAN connectivity
4.8 Tbps	LCG Core

CINBAD codename deciphered

CERN **I**nvestigation of **N**etwork **B**ehaviour and **A**nomaly **D**etection

Project Goal

"T*o* ***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 "*

What is anomaly? (1)

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

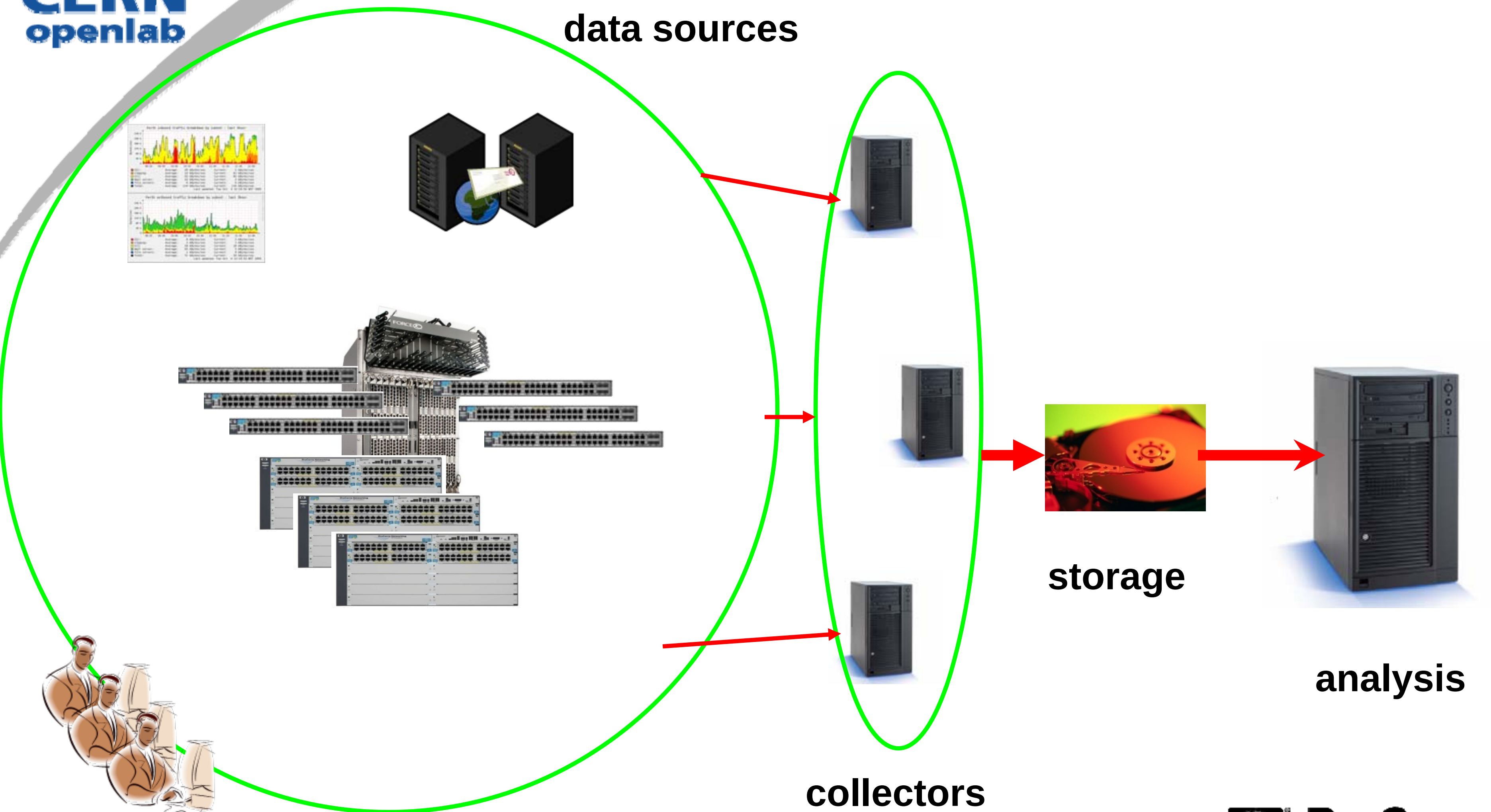
Network faults	Malicious attacks	Viruses/worms
Misconfiguration	…	…

- 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"*

What is anomaly? (2)

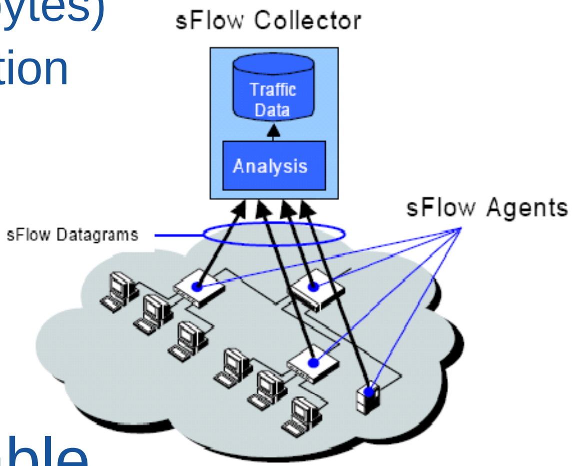
- Just a few examples of anomalies:
  - The network infrastructure misuse
    - unauthorised DHCP/DNS server (either malicious or accidental)
    - network scans
    - worms and viruses
  - Violation of a local network/security policy
    - NAT, TOR usage (not allowed at CERN)

CINBAD project principle



sFlow – main data source

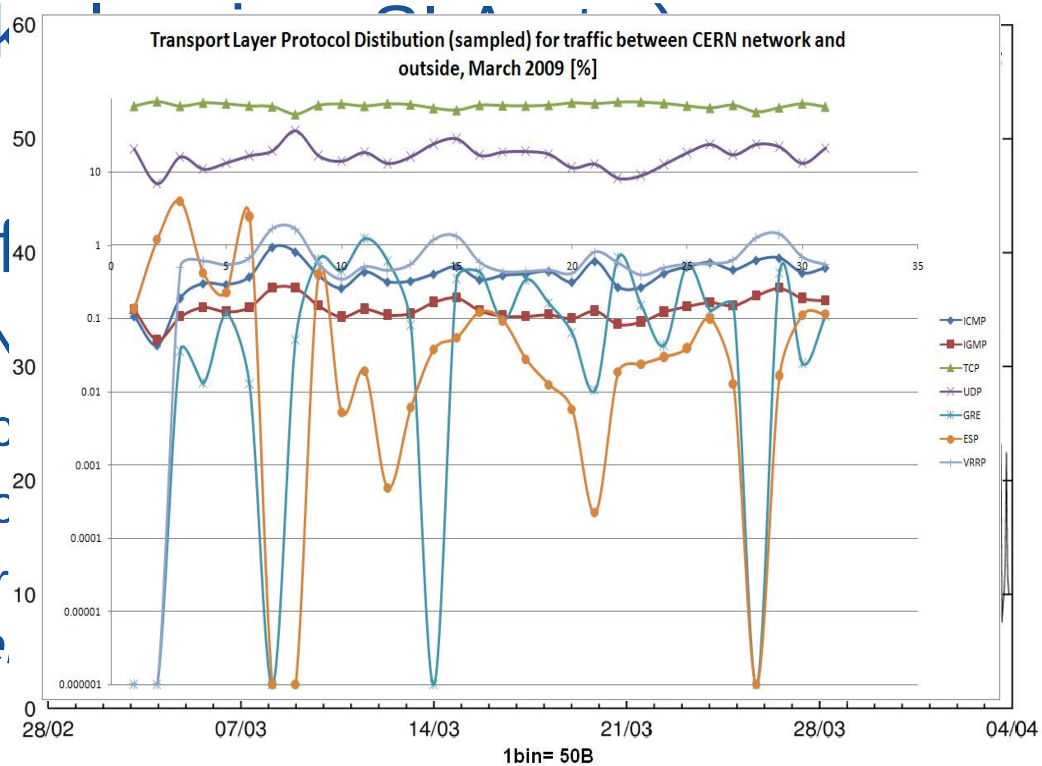
- Based on packet sampling (RFC 3176)
  - on average 1-out-of-N packet is sampled by an agent and sent to a collector
    - packet header and payload included (max 128 bytes)
      - switching/routing/transport protocol information
      - application protocol data (e.g. http, dns)
  - SNMP counters included
  - low CPU/memory requirements – scalable



- For more details, see our technical report

http://openlab-mu-internal.web.cern.ch/openlab-mu-internal/Documents/2_Technical_Documents/Technical_Reports/2007/RJ-MM_SamplingReport.pdf

- Typically: traffic accounting (e.g. for billing, networ



- Useful f … s
  - CERN
    - acc … ffic
    - info … seen
    - par … *uld not be the*

- Rare examples of sflow usage for anomaly detection

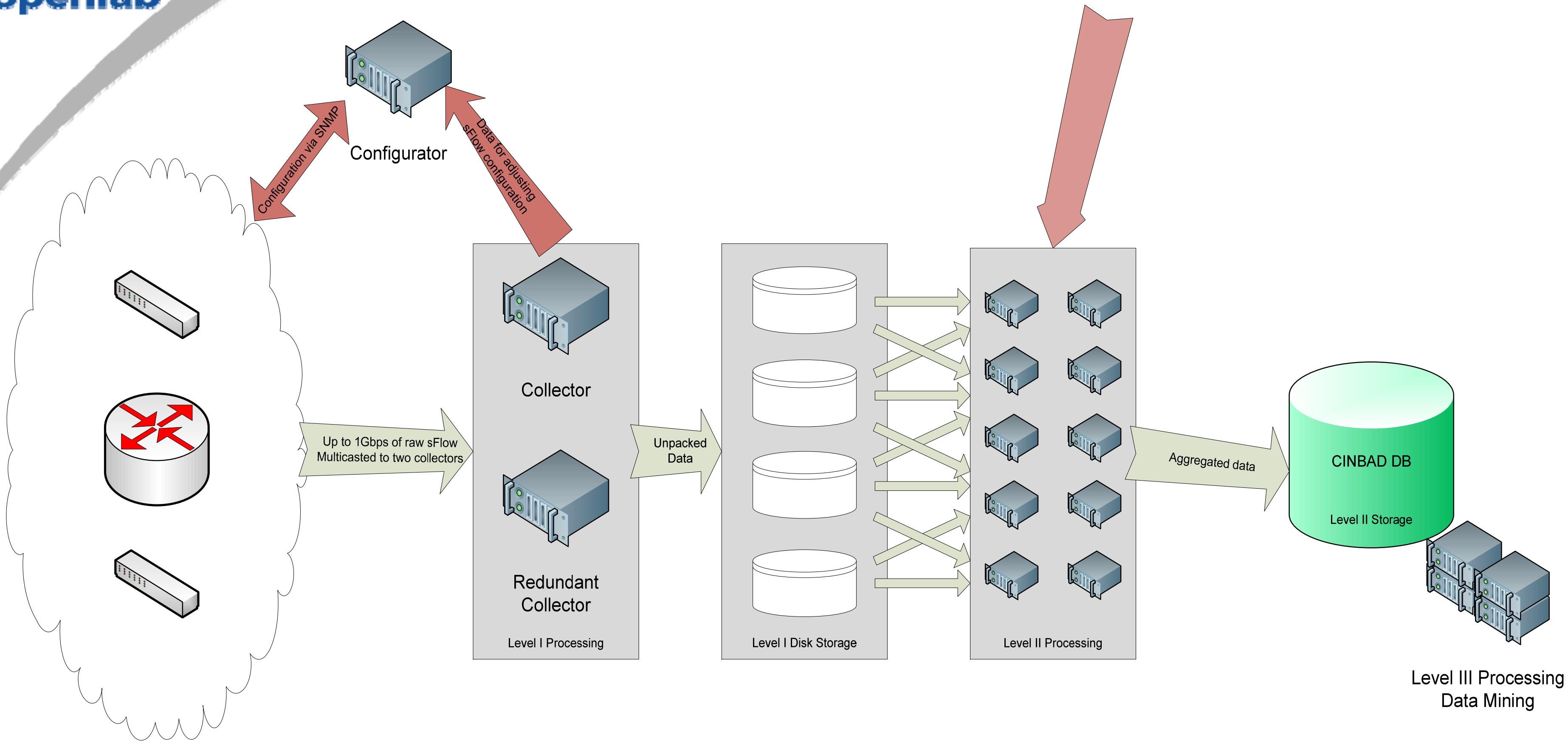
Other data sources

- Packet sampling data is not enough!
  - Data is partial, cannot provide 100% accuracy
  - Not always easy to identify the anomaly

- More data to understand flow of data in the network
  - External sources provide useful information and time triggers
  - Correlation between various data sources

- Example:
  - Central Antivirus Service at CERN

CINBAD sFlow data collection

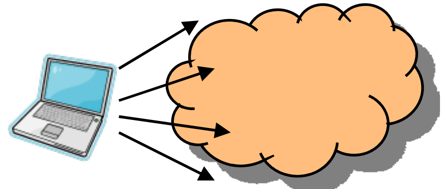


- ✓ Current collection based on traffic from ~1000 switches
  - ✓ ~6000 sampled packets per second
  - ✓~ 3500 snmp counter sets per second

Multistage data storage

- Stage 1
  - sFlow datagram tree-like format unpacked into CINBAD file format to enable fast direct access
  - Minimal space overhead introduced

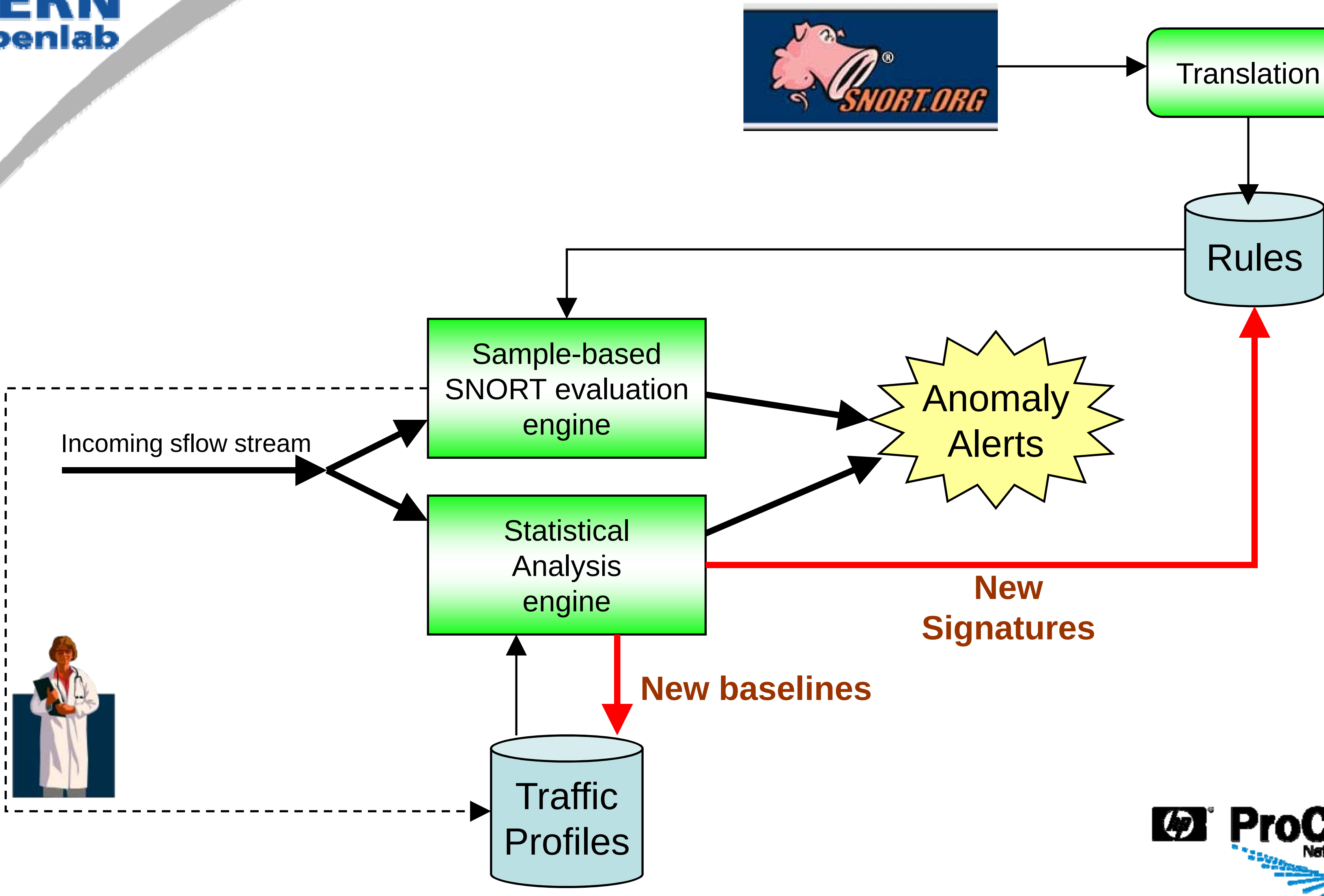
- Stage 2
  - Oracle DB as a long-term storage
  - data aggregation
    - tradeoff between sFlow randomness and space, data lifetime and anomaly detection requirements
      – e.g. number of destination IPs for a given source IP



Data analysis

- Various approaches are being investigated
  - Statistical analysis methods
    - detect a change from “normal network behaviour”
      - selection of suitable metrics is needed
    - can detect new, unknown anomalies
    - poor anomaly type identification
  - Signature based
    - **we ported SNORT to work with sampled data**
    - performs well against know problems
    - tends to have low false positive rate
    - does not work against unknown anomalies

Synergy from both detection techniques



Current results

- Campus and Internet traffic analysis
  - Identified anomalies which went undetected by the central CERN IDS
  - Detected number of misbehaviors
    - both statistical and pattern matching approaches used
    - TOR, DNS abuse, Trojans, worms, network scans, p2p applications, rogue DHCP servers, etc.
- Findings reported to the CERN security team
  - Security team adapted their policies

Achievements and next steps

- It has been demonstrated that pattern matching for anomaly detection is possible with sflow data
  - **Payload data is a key advantage of sflow**
  - sflow allows distributed detection at the edge of the network
- Combining statistical analysis with pattern matching is providing encouraging initial results
- Integration of the two mechanisms holds promise for zero-day detection