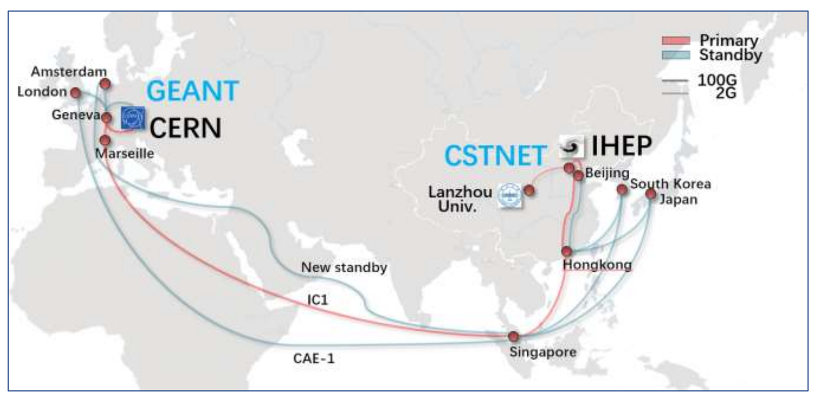
Research of Wide Area Network Performance Anomaly Detection Technology Based on Machine Learning

Shan Zeng[1,2], Cheng Li[1,2] on behalf of IHEP network group
[1]Institute of High Energy Physics
[2]University of Chinese Academy of Sciences
zengshan@ihep.ac.cn

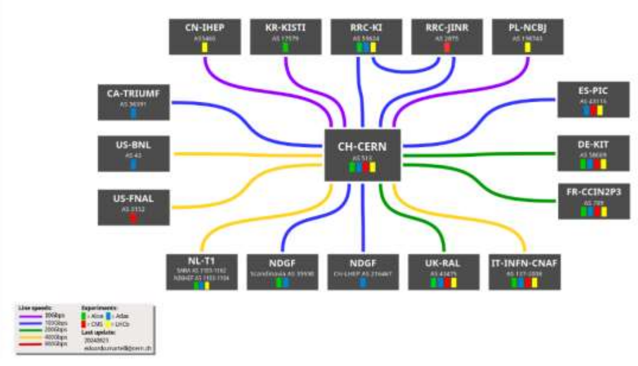
Introduction

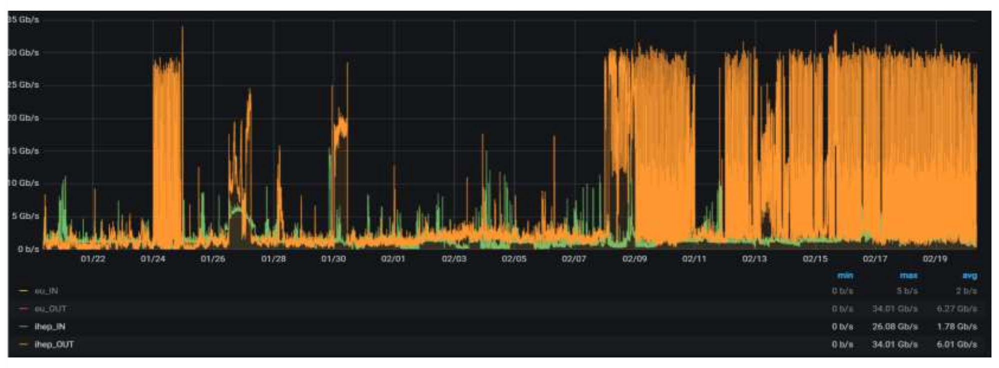
IHEP endorsed as a new WLCG Tier-1 site (June,2024), WAN bandwidth was upgraded from 40Gbps to 100Gbp

- LHCOPN@IHEP
  - 20Gbps bandwidth guaranteed
  - 3 links redundancy
  - ~ 200ms latency
- LHCONE@IHEP
  - 100Gbps bandwidth shared



LHCOPN





Many network challenges from daily network operation

- Issue debugging is difficult and time-consuming
- How to thoroughly and vividly demonstrate various network measurement results to the application
- How to promptly detect and resolve the network issues

Network performance R&D is essential in view of HL-LHC

- Effective network usage and prompt detection as well as resolution of any network issues need to be guaranteed

Reports from CHEP/HEPiX/LHCOPN-LHCONE meeting

- *Shawn*: Analyzing, Identifying & Alerting on Network Issues[1]
- *Petya*: perfSONAR Network Analytics through Machine Learning[2]

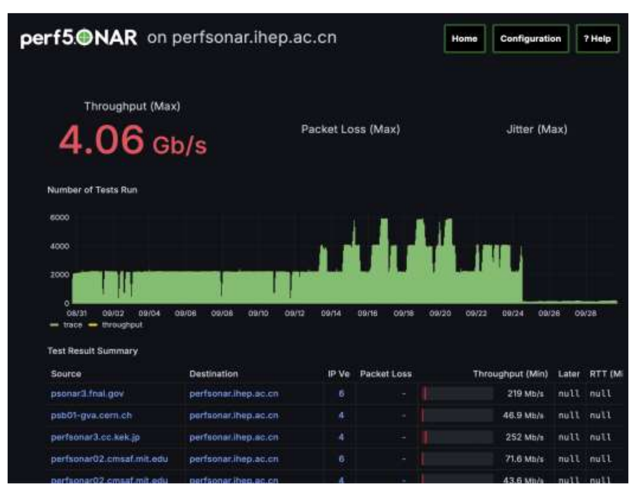
The network performance needs to be closely monitored and evaluated
Network analytics R&D is essential for providing high quality network services
Machine learning methods seem well-suited to solving these types of problems

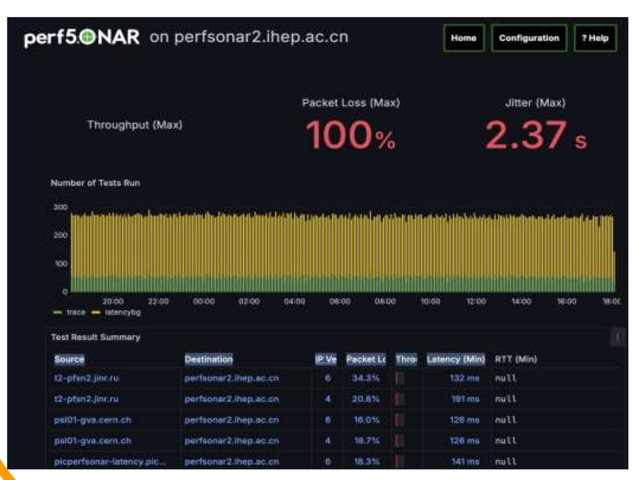
Related works

Active measurement of network performance

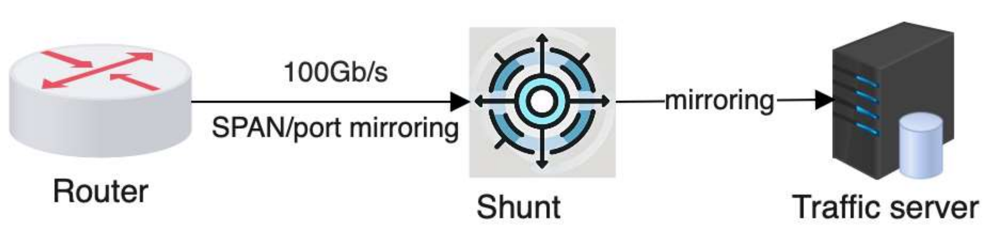
- IHEP perfSONAR upgraded to the latest version: v5.1.3

IHEP WAN traffic are captured and stored in local file system





- Full traffic packet captured, in case of issue omitted
  - Captured by tcpdump, stored as .cap file
  - every 10 minutes a file, data volume is 1.4TB-7TB per day
- In-depth understanding of the network communication
  - Establish connection, data transmission, release connection …
- Find out the root cause of problems during communication between applications



Architecture design

What we get?

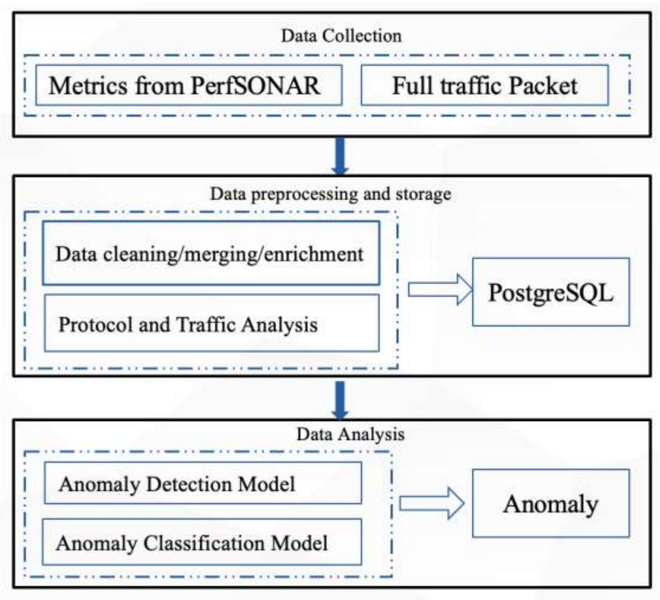
- WAN performance monitoring metrics from perfSONAR
- WAN full traffic packet by mirroring

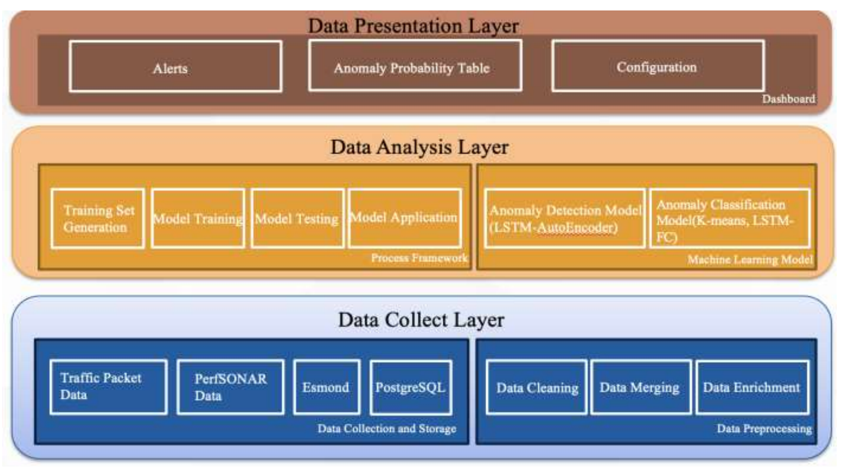
What we want?

- Find network anomalies when exist
- highlight the time periods of these anomalies
- provide a classification table of anomaly types
- Identify the anomaly classification and the time it occurs

How we did?

- Data cleaning to remove invalid data
- Data merging to merge perfSONAR metrics and traffic packet
- Data enrichment to enrich the institute name and its nodes
- PostgreSQL for storage
- ML model for analyzing
  - Anomaly detection
  - Anomaly classification





Data Collect Layer

- Collect perfSONAR metrics data through Esmond API
- Analyze the JSON data return from Esmond, after data cleaning, merge with the traffic packet data
- Enriching the data with institution information
- Install them in the data warehouse: PostgreSQL

Data Analysis Layer

- Anomaly detection model(based on LSTM-AutoEncoder)
- Anomaly detection model(based on K-means&LSTM-FC)

Data Presentation Layer

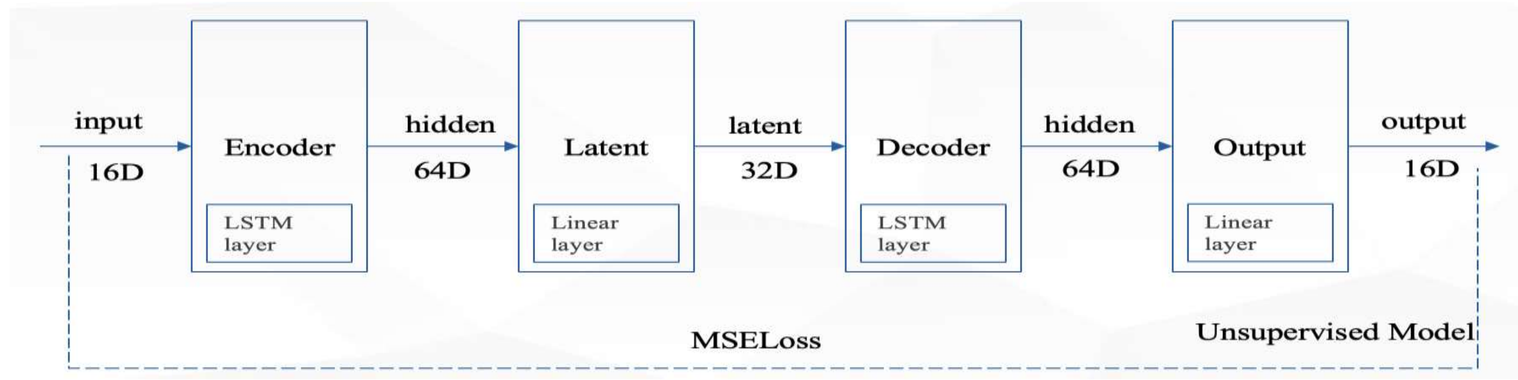
- Provide interface to other systems/platforms
- Provide configuration dashboard to administrators

Anomaly detection model

LSTM autoencoder model was designed

- The reconstruction loss is first computed using the autoencoder. If the reconstruction loss is large, the data is considered to be anomalous

Column	Type
timestamp_unix	bigint
src_ip	text
dst_ip	text
src_port	integer
dst_port	integer
perfsonar_src_ip	text
perfsonar_dst_ip	text
protocol	text
packet_size	integer
ttl	integer
dscp	integer
window_size	integer
bandwidth_utilization	double precision
latency	numeric[]
packet_loss_rate	double precision
src_organization_domain	text
dst_organization_domain	text
anomaly_type	text



Encoder: The LSTM processes information at each time step and encodes it into a 64-dimensional space

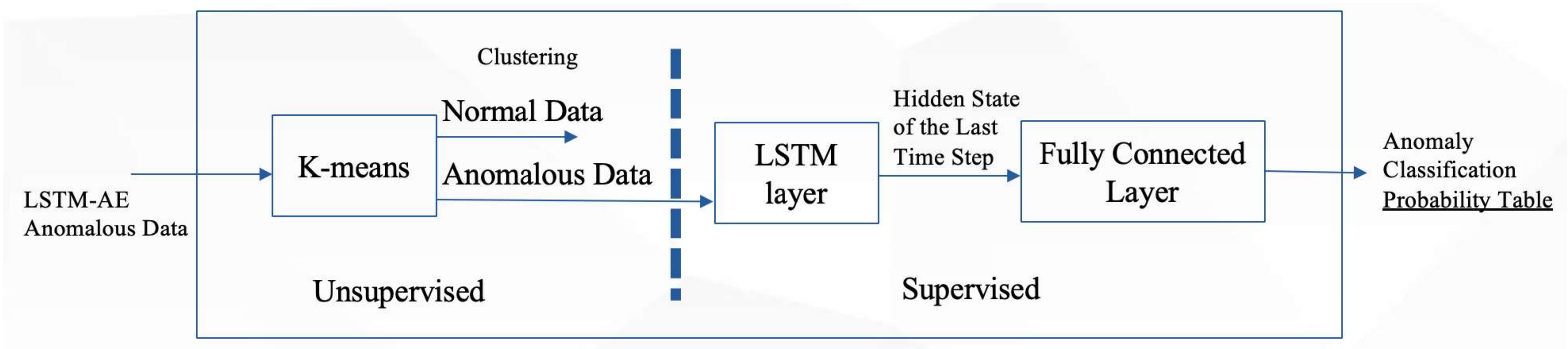
Decoder: This layer decodes the latent vector back to the input sequence shape

Latent Layer: It compresses the hidden state into a lower-dimensional latent vector

Output Layer: It transforms the decoder's hidden state into a reconstruction sequence matching the original input dimensions

Anomaly classification model

K-means and LSTM model was designed



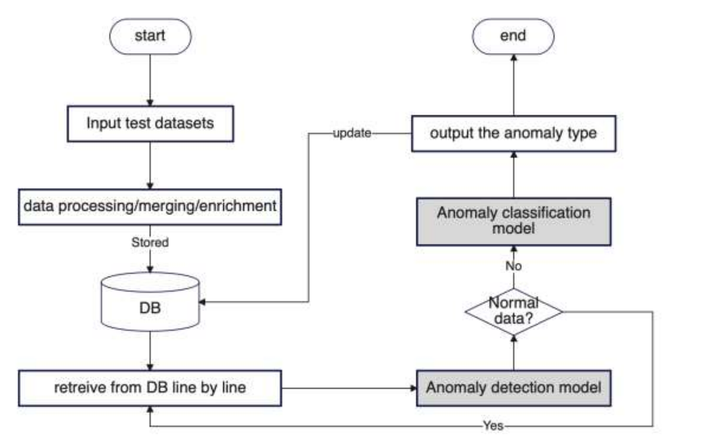
- To identify previously undiscovered types of anomalies, the K-means algorithm is used to cluster the anomalous data
- A fully connected layer is utilized to determine the specific categories of the anomalous data

Workflow

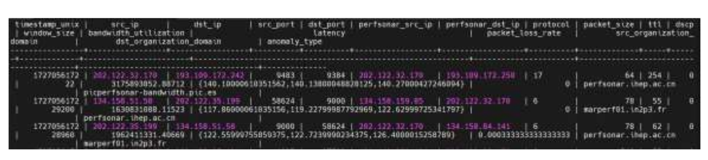
Step1: Train the ML models using the training dataset

Step2: Tuning the model parameters to make sure the ML model is ready

Step3: Testing started…



- **Done**: Table created, metrics inserted via Python, parallel processes for dataset creation, LSTM-AE anomaly detection model developed
- **To Do**: Fix missing metrics, increase test datasets, enhance dataset size, improve processing efficiency, strengthen LSTM-AE for missing data, boost model resilience, develop anomaly classification, design alerts



references

[1] Vasileva, Petya, et al. "Analyzing, Identifying & Alerting on Network Issues." EPJ Web of Conferences, vol. 295, 2024, p. 07003. EDP Sciences, https://doi.org/10.1051/epjconf/202429507003.
[2] Vasileva, Petya, et al. "perfSONAR Network Analytics: Status & Plans." CERN, 2024.