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

Shan Zeng, Cheng Li
on behalf of IHEP network group
Funded by NSFC (No. 12175258)

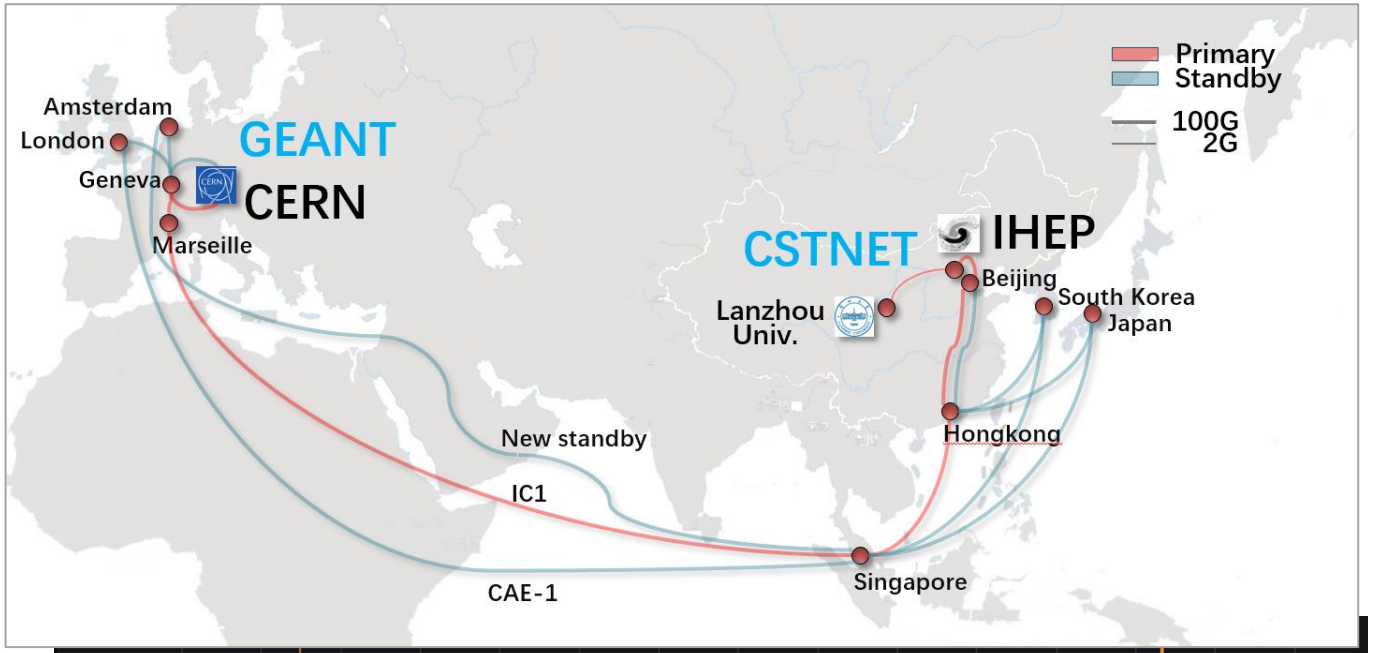
zengshan@ihep.ac.cn
2024/10/10

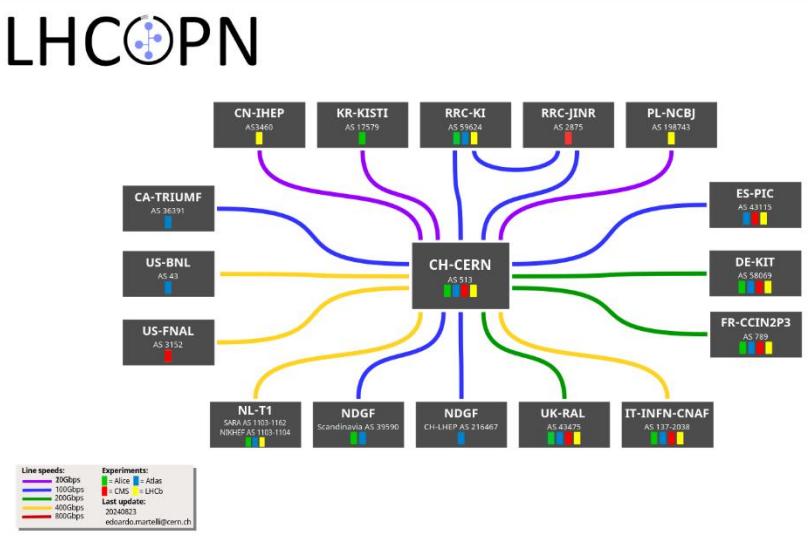
Outline

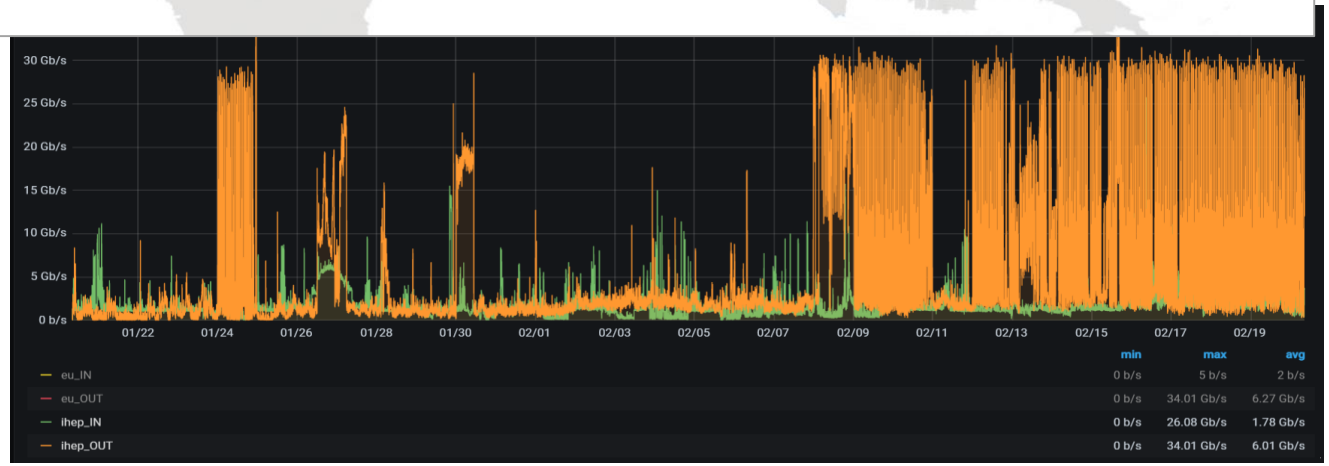
- Background
- Related works
- Architecture design
- Analysis method and process
- Research progress
- Future plan
- Summary

Background

- **IHEP endorsed as a new WLCG Tier-1 site (June,2024), WAN bandwidth was upgraded from 40Gbps to 100Gbps**
  - LHCOPN@IHEP
    - 20Gbps bandwidth guaranteed
    - 3 links redundancy
    - ~ 200ms latency
  - LHCONE@IHEP
    - 100Gbps bandwidth shared







Background： network challenges

- **Network is a critical part of WLCG's infrastructure, becomes more and more important to assure the site availability and reliability**
- **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

Background： current status of peer research

- **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
    - *https://indico.jlab.org/event/459/contributions/11662/attachments/9322/13521/CHEP-Poster-NetAnalytics-Final.pdf*
  - *Petya:* perfSONAR Network Analytics through Machine Learning
    - *https://indico.cern.ch/event/1410638/contributions/6127645/attachments/2944638/5174511/perfSONAR%20Network%20Analytics%20-%20Status%20&%20Plans.pdf*

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**

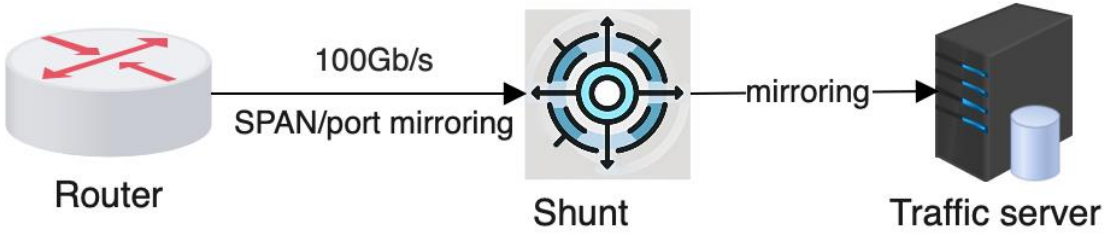
  – Full traffic packet captured, in case of issue omitted

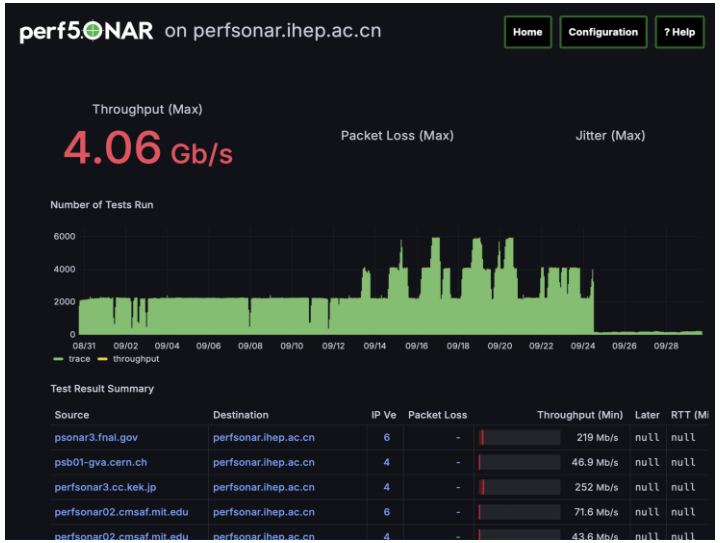
    - Captured by tcpdump, stored as .cap file
    - every 10 minutes a file, data volume is 1.4TB-7TB per day

  – in-depth understanding of the network communication

    - Establish connection, data transmission, release connection …

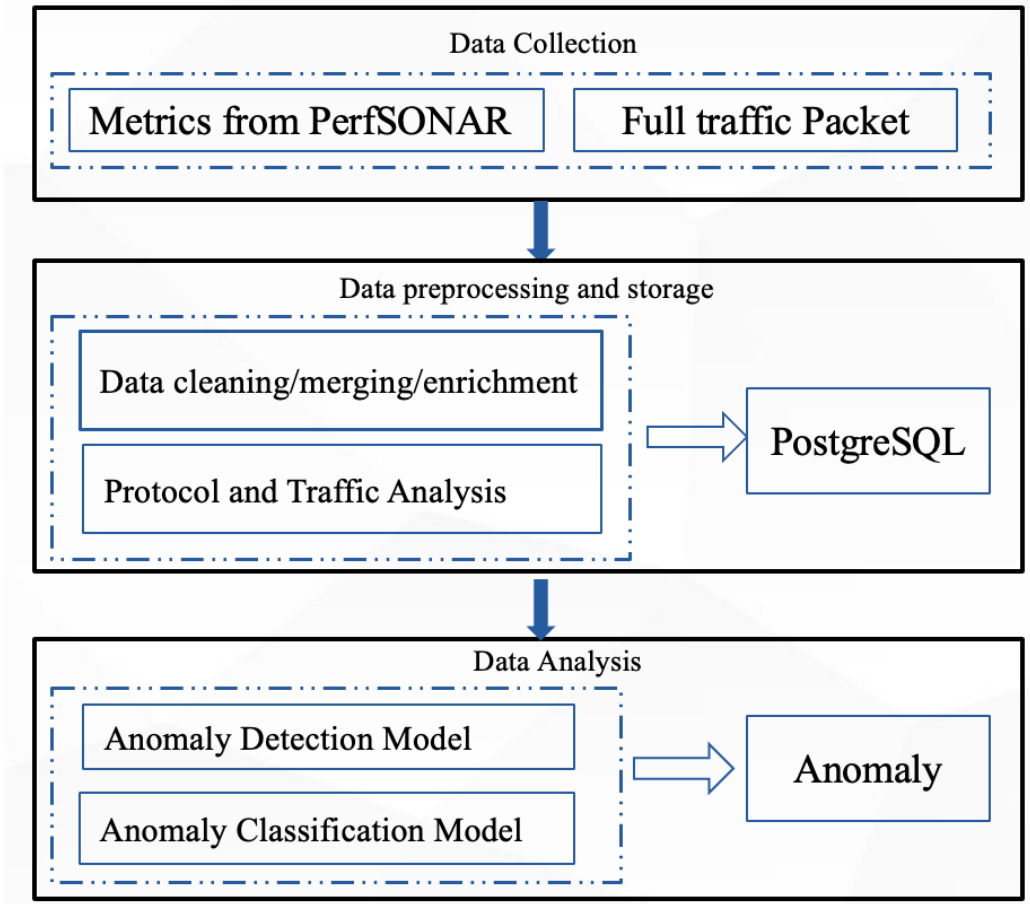
  – Find out the root cause of problems during communication between applications





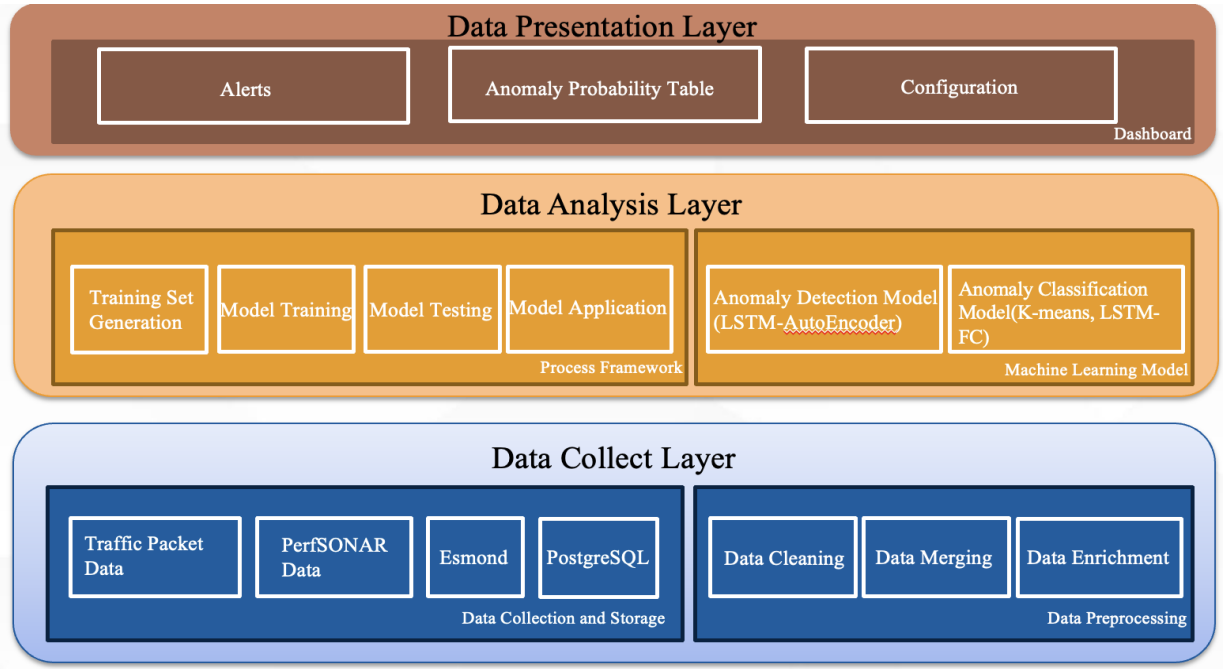
Architecture design

- **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



3-layer structure design

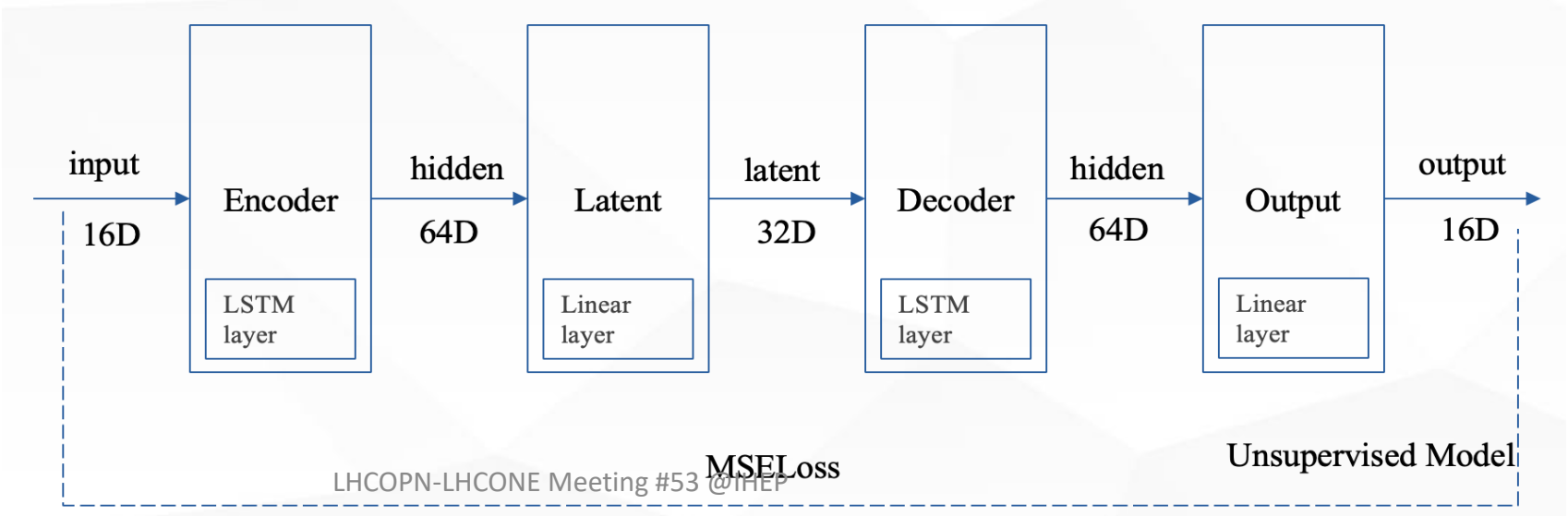
- **Data Collect Layer**
  - Collect perfSONAR metrics data through Esmond API
  - Analyze the JSON data return from Esmond, after data cleaning, merge with the traffic packet data
  - Enriching the data with institution information
  - Install them in the data warehouse: PostgreSQL
- **Data Analysis Layer:** two ML models are provided
  - Anomaly detection model
    - based on LSTM-AutoEncoder
  - Anomaly classification model
    - based on K-means&LSTM-FC
- **Data Presentation Layer**
  - Provide interface to other systems/platforms
  - Provide configuration dashboard to administrators



Anomaly detection model

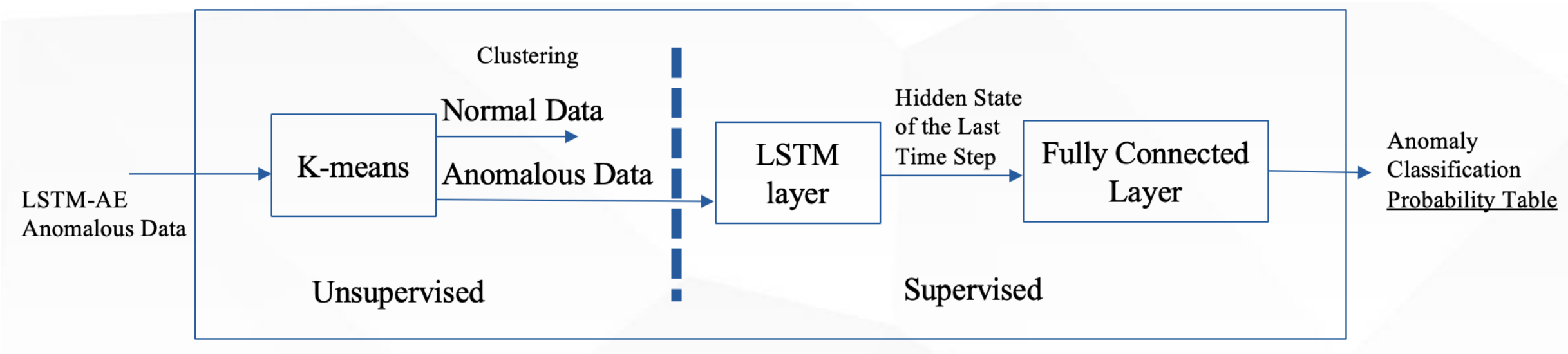
- **LSTM autoencoder model was designed**
  - the reconstruction loss is first computed using the autoencoder. If the reconstruction loss is large, the data is considered to be anomalous
  - Encoder: LSTM extracts information at each time step and stores it in a 64-dimensional space
  - Latent layer: Extract the hidden state and compress it into a lower-dimensional latent vector
    - The dimensionality reduction process can be viewed as 'compressing' complex high-dimensional data and extracting the most important and informative features.
  - Decoder: Decode the latent vector back to the shape of the input sequence
  - Output layer: Convert the hidden state of the decoder LSTM back into a reconstruction sequence with the same dimension as the original input.

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



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

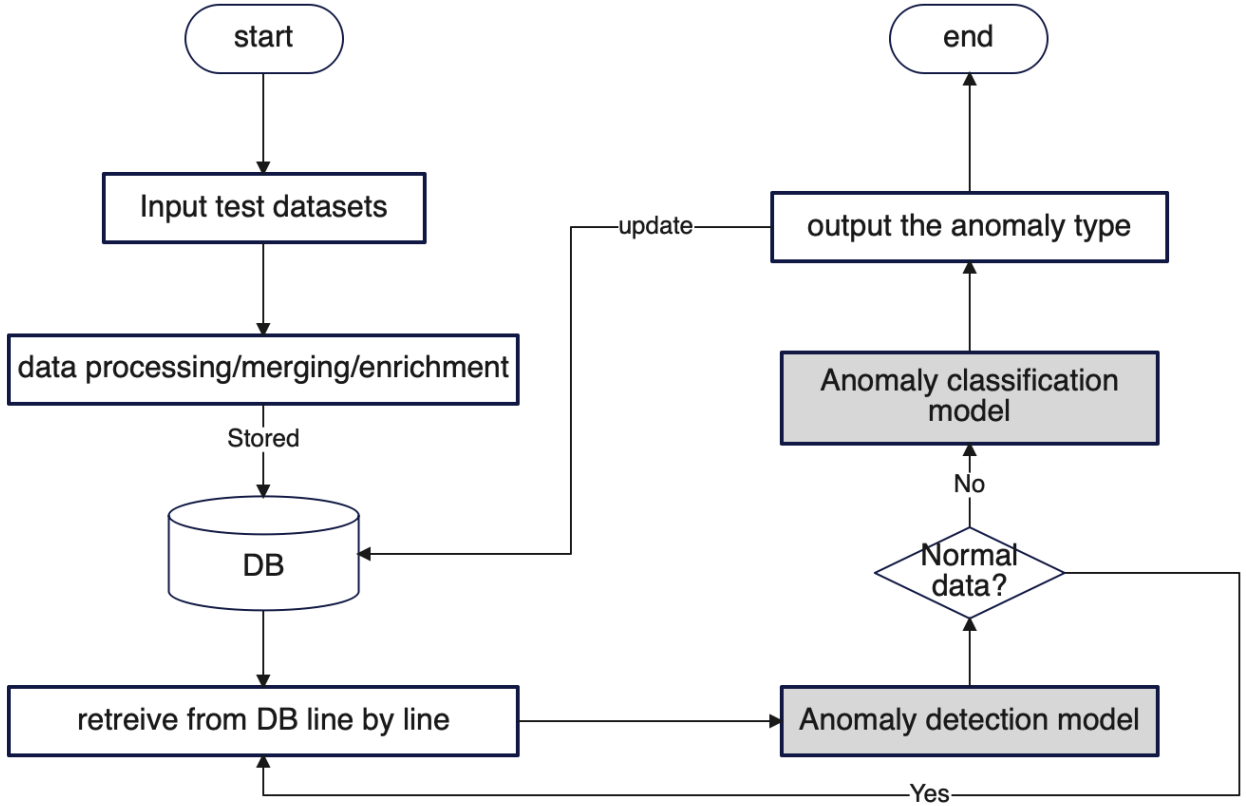


How it works

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…



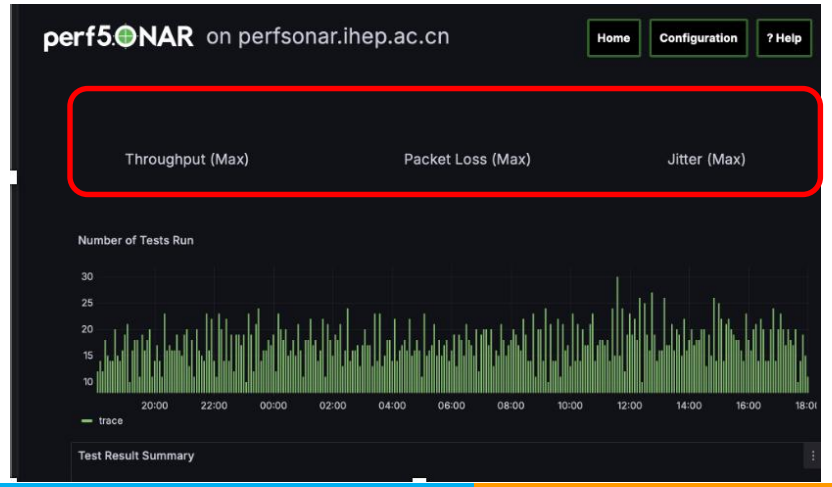
Research progress

Data engineering

- Table has been created and metrics data can be inserted automatically through Python scripts
- parallel processes are running background to provide high performance automatic datasets creation
- Issues we meet: some metrics data from the newest perfSONAR are missing

Training model

- The unsupervised training model for the LSTM-AE anomaly detection part has been developed
  - using normal data to train the LSTM-AE model
  - However, due to insufficient dataset size and parameter tuning issues, the training results aren't ideal
  - Efforts should be done to overcome these challenges.




```
 timestamp_unix |     src_ip      |     dst_ip      | src_port | dst_port | perfsonar_src_ip | perfsonar_dst_ip | protocol | packet_size | ttl | dscp
 | window_size | bandwidth_utilization |                                  latency                                  |   packet_loss_rate   |        src_organization_
domain        |        dst_organization_domain        | anomaly_type
---------------+-----------------+-----------------+----------+----------+------------------+------------------+----------+-------------+-----+-----
-+-------------+-----------------------+-------------------------------------------------------------------------+----------------------+--------------------------
-------------+---------------------------------------+--------------
     1727056172 | 202.122.32.170  | 193.109.172.242 |     9483 |     9384 | 202.122.32.170   | 193.109.172.250  | 17       |          64 | 254 |    0
 |          22 |      3175893852.88712 | {140.10000610351562,140.13800048828125,140.27000427246094} |                    0 | perfsonar.ihep.ac.cn
             | picperfsonar-bandwidth.pic.es         |
     1727056172 | 134.158.51.58   | 202.122.35.199  |    58624 |     9000 | 134.158.159.85   | 202.122.32.170   | 6        |          78 |  55 |    0
 |       29200 |      1630831088.11523 | {117.86000061035156,119.22799987792969,122.62999725341797} |                    0 | marperf01.in2p3.fr
             | perfsonar.ihep.ac.cn                  |
     1727056172 | 202.122.35.199  | 134.158.51.58   |     9000 |    58624 | 202.122.32.170   | 134.158.84.141   | 6        |          78 |  62 |    0
 |       28960 |      1962411331.40669 | {122.55999755859375,122.7239990234375,126.4000015258789}   | 0.000333333333333333 | perfsonar.ihep.ac.cn
             | marperf01.in2p3.fr                    |
```

Future plan

- **ElasticSearch/OpenSearch cluster is considered to be used to handle the huge amount of data sets**
- **Increase the quantity of the test data sets**
  - Enhance the dataset size to provide more comprehensive testing
  - Processing efficiency of data engineering should also be concerned
- **Strengthen the LSTM-AE model's ability to handle missing data**
  - Focus on adequately training the model to improve its resilience to data gaps
- **Develop the anomaly classification model**
  - Once the anomaly detection component is completed, proceed to design the model for anomaly classification
- **Design and develop alert visualizations**

Summary

- **The purpose is to quickly find the network anomalies through network performance assessment**
  - Based on the newest version of perfSONAR and full network traffic packet analysis
- **We started to do the research since the middle of this year**
  - Architecture design was finished
  - Recently most work focused on data engineering
  - More exchanges with the perfSONAR team will be conducted
  - Some functions of analysis model have been developed
  - More functions need to be developed and optimized
- **Any suggestions and cooperation are welcomed and needed**

Thanks for your attentions

Questions, Comments, Suggestions?