CERN openlab

ALICE Grid throughput prediction

Costin Grigoras, Mircea-Marian Popa, **Sofia Vallecorsa**

CERN

Sofia.Vallecorsa@cern.ch

29/11/2021

Outline

Model the MonALISA I/O throughput using monitoring data

- Focus on **READ queries (READ/WRITE ratio is 30:1)**

Study a set of RNN-based architectures

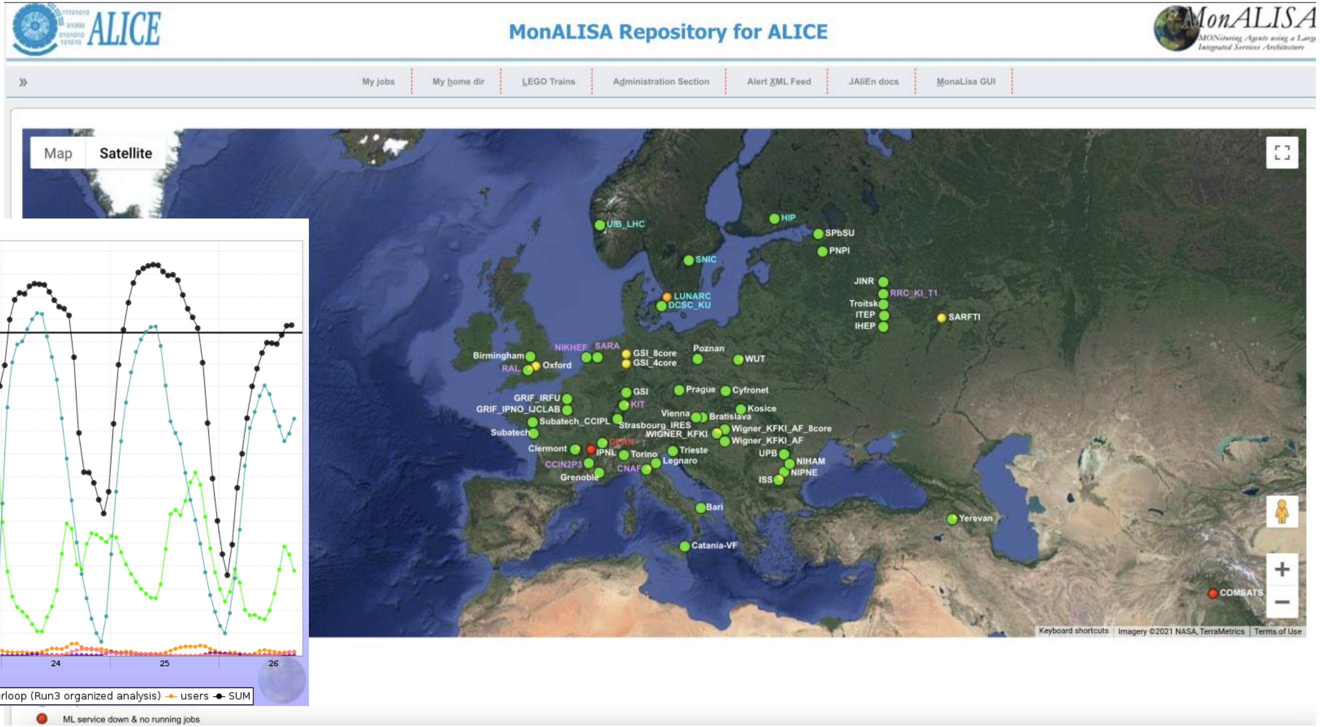
- Throughput modeling as a function of network topology (project funded by ATTRACT)
  - PCA + LSTM
  - Auto-Encoder + LSTM
- Throughput forecasting through end-to-end Seq2Seq inspired model

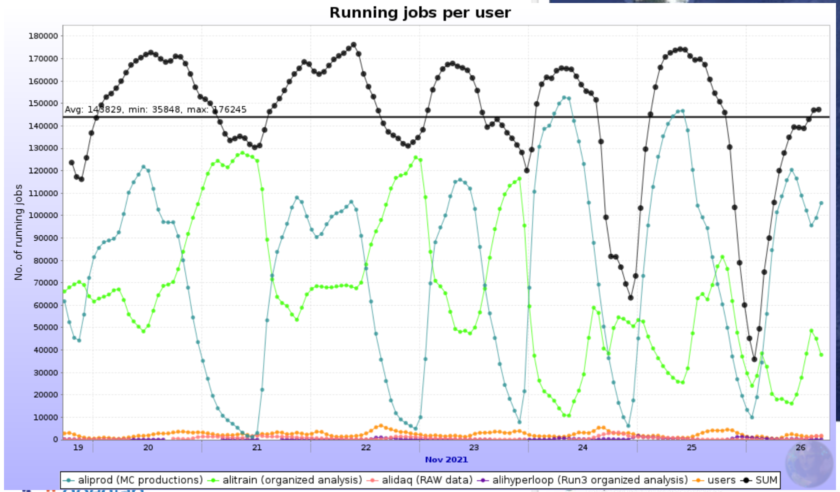
http://alimonitor.cern.ch/

MonALISA

Legrand I., *et al.*, Computer Physics Communications 180 (2009) 2472–2498

An agent based, dynamic service system to monitor, control and optimize distributed systems

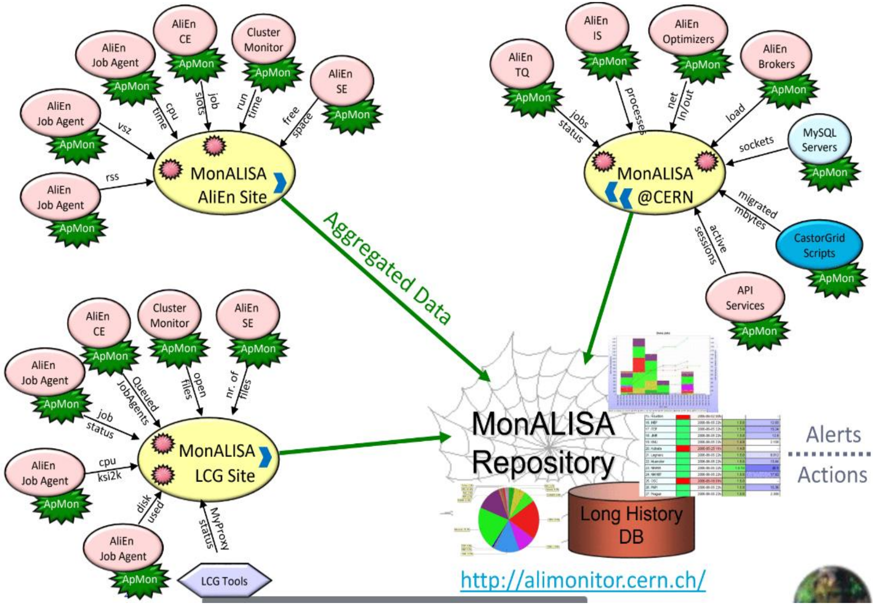




Automatic storage discovery

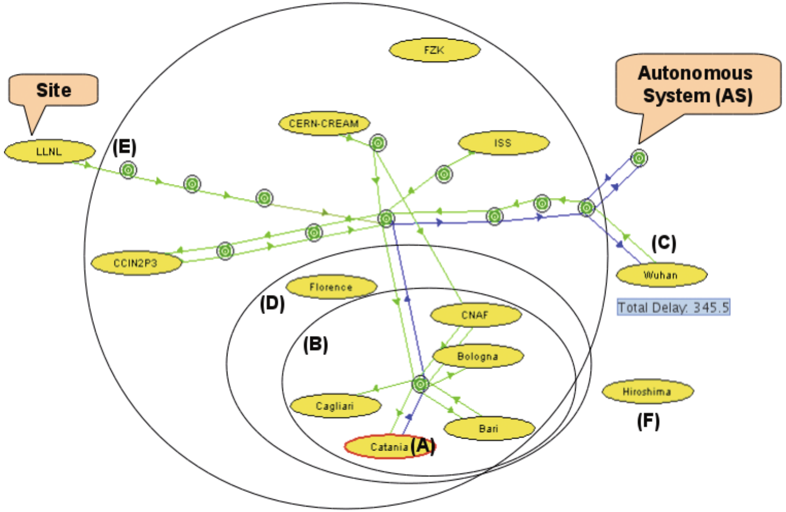
Selection of storage elements for file access

- Answers **2 types of questions (queries)** from clients:
  - which is the optimal SE to **READ** a input file from ?
  - where to **WRITE** output data ?
- Gathers data on **network topology** and **SE usage and functionality**
- A set of **central nodes** answer queries using a **heuristic method**



Heuristic network topology discovery

- Compute client-to-storage **distance**:
  - **distance** matrix aggregates data on network topology (~6700 elements)
  - **demotion** matrix describes SE availability
- Build sorted SE list
- MonALISA logs monitoring data using a 2 minutes time step
  - Distance and demotion matrices are updated asynchronously



Input data pre-processing (I)

Queries Readsize

- List of queries, each containing
  - Computing farm
  - File read size (RS)
  - SEs containing a copy of the file
- A file will most likely be read from CERN SE when it is the first option

Time Stamp	Read Size (Bytes)	Computing Farm	SE list
17000	12360	CERN	IHEP; CCIN2P3; CERN
19000	102360	CCIN2P3	CEPH; CCIN2P3; CERN
17000	34360	CCIN2P3	RAL; SINC; CERN

Sort entries according to distance and demotion matrices

Time Stamp	Read Size (Bytes)	Computing Farm	SE list
17000	12360	CERN	CERN; CCIN2P3; IHEP
19000	102360	CCIN2P3	CCIN2P3; CERN; CEPH
17000	34360	CCIN2P3	CERN; RAL; SINC

Select queries were CERN SE is the first option.

Time Stamp	Read Size (Bytes)	Computing Farm	SE list
17000	12360	CERN	CERN; CCIN2P3; IHEP
17000	34360	CCIN2P3	CERN; RAL; SINC

Drop computig Farm and SE list

Time Stamp	Read Size (Bytes)
17000	12360
17000	34360

Sum entries with same time stamp

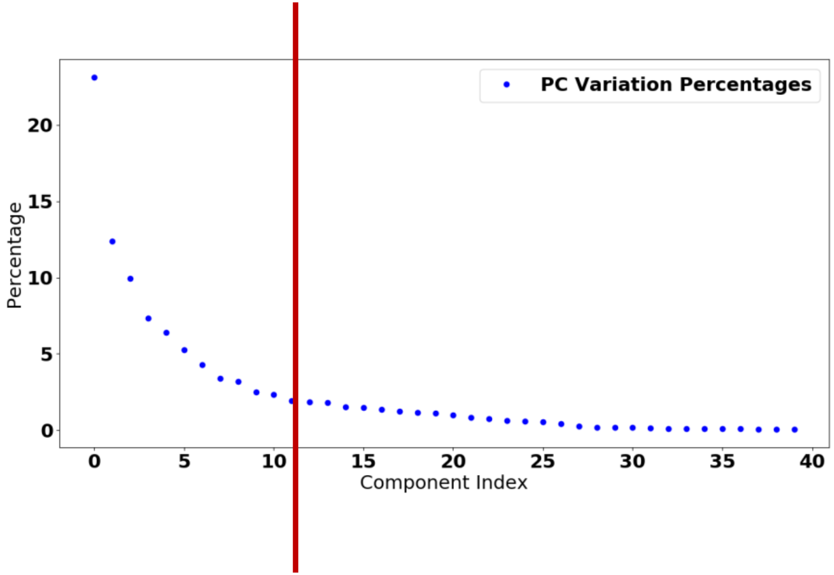
Time Stamp	Read Size (Bytes)
17000	**46720**

CERN openlab

Input data pre-processing (II)

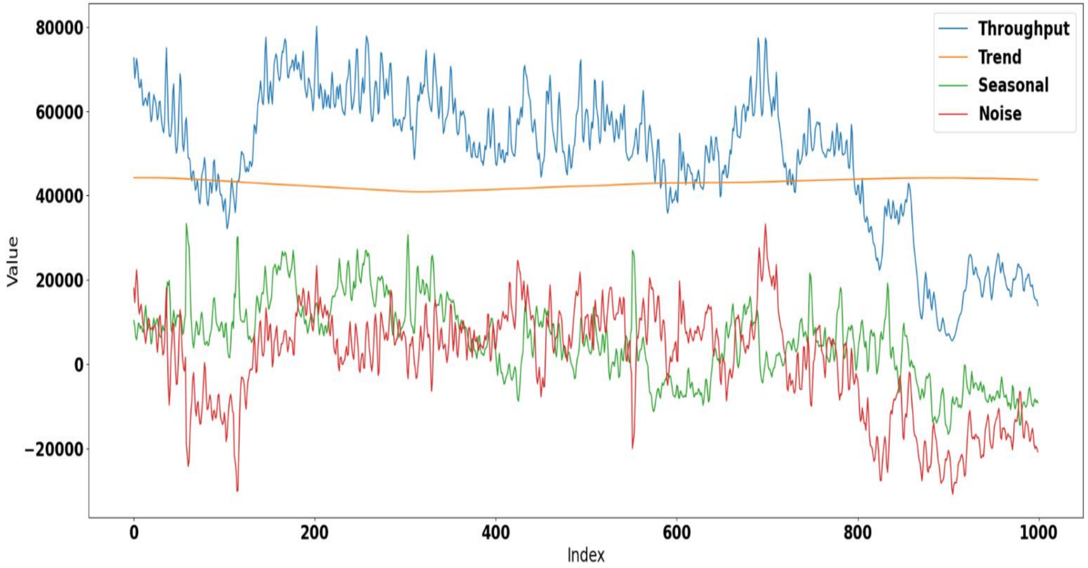
Distance matrices dimensionality reduction

- Predict throughput taking into account the network status as monitored by MonALISA through the demotion and distance matrices
- Reduce dimensionality by:
  - Performing a principal component analysis
    - used **11 components** which accounted for ~95% of the variance
  - Using auto-encoder neural networks to project to a **10-dimensional latent space**



MonALISA throughput

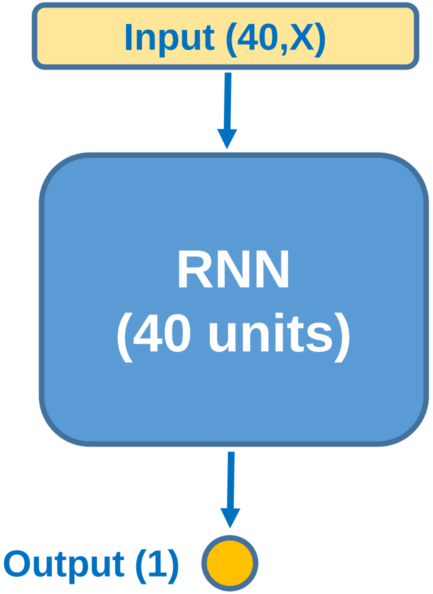
- Large fluctuations in the observed network utilization
- Time series decomposition
  - Model noise using 0-centered Gaussian distribution at a frequency of ~**5.91 days**
- Use the time stamp to associate throughput to input data

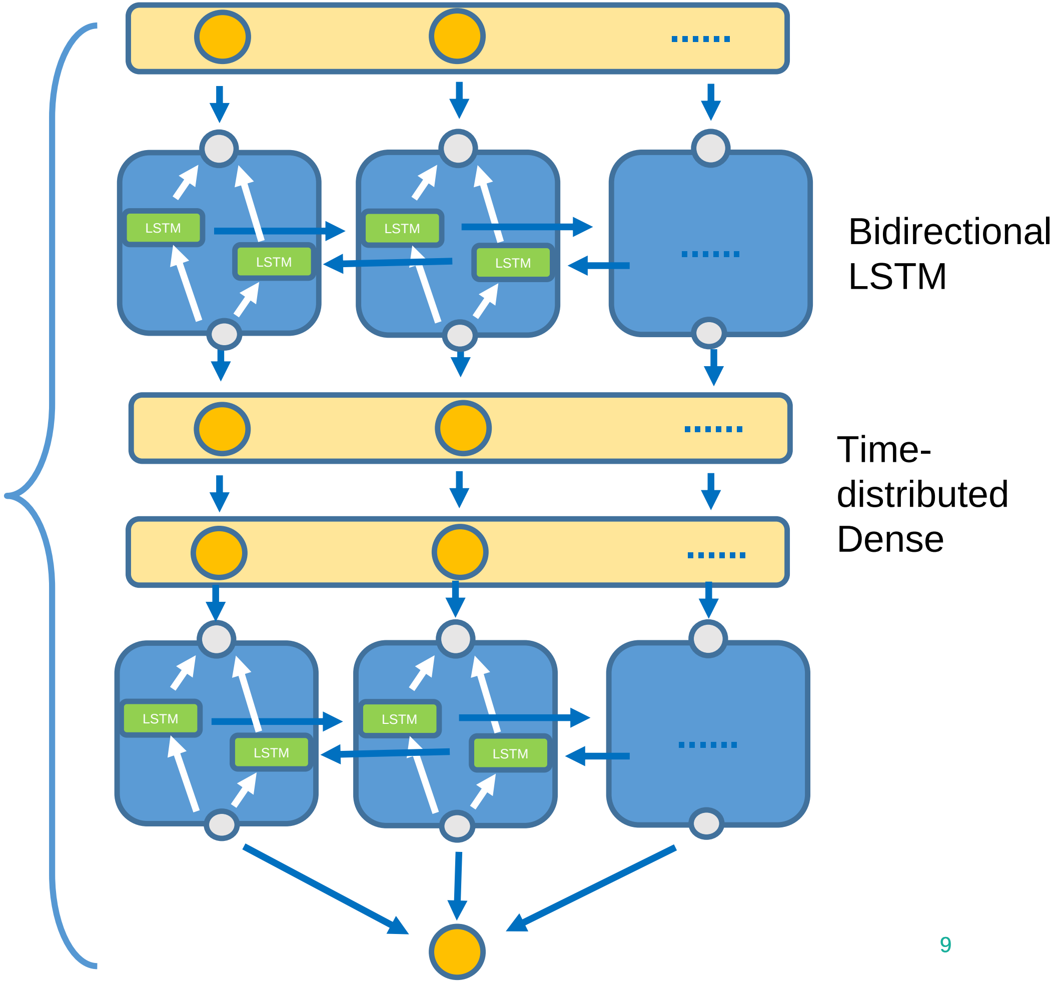


Training dataset:
17 January 2020 -> 24 January 2020
31 January 2020 -> 9 February 2020
9 February 2020 -> 14 February 2020

Validation dataset:
13 May 2020 -> 27 May 2020

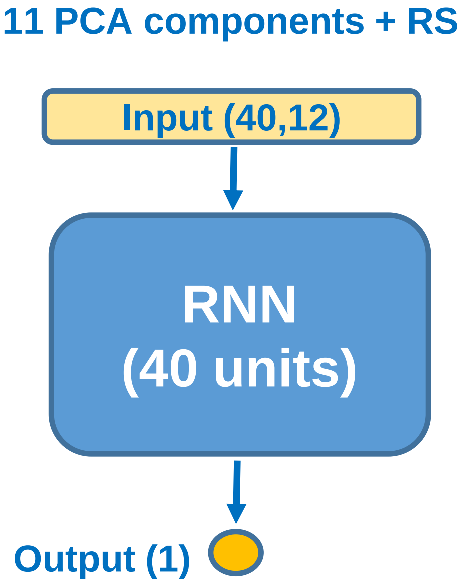
RNN architecture



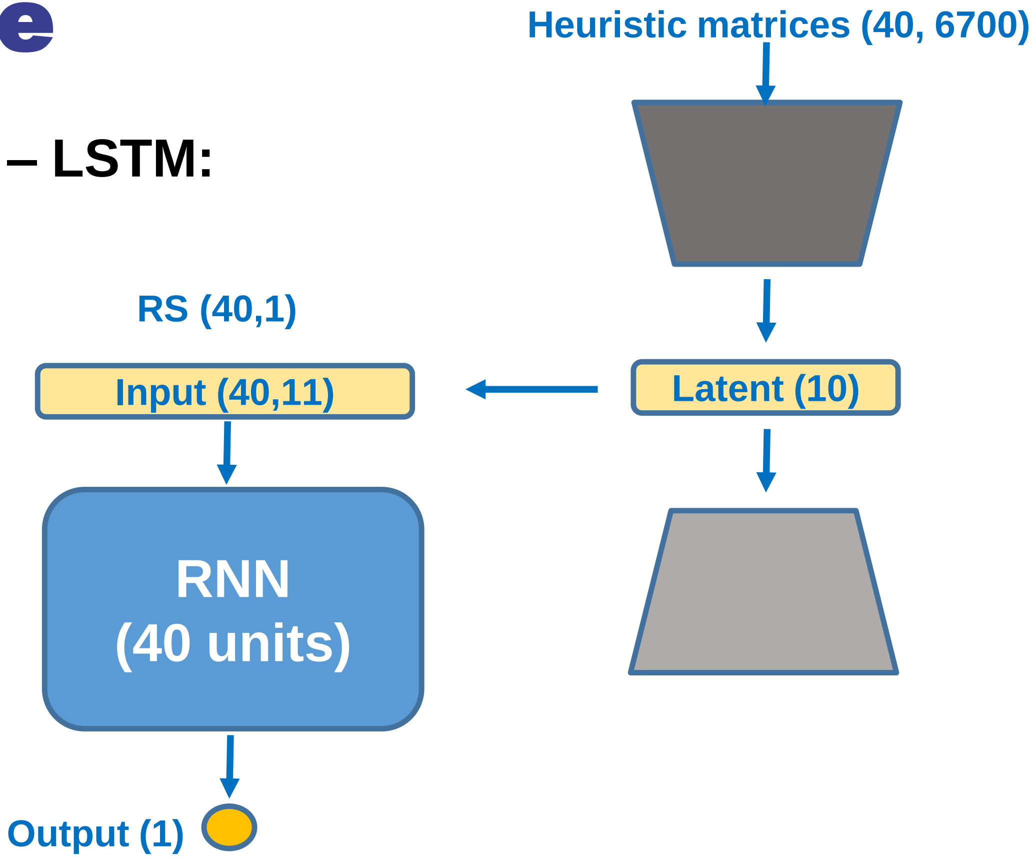


RNN architecture

PCA – LSTM:



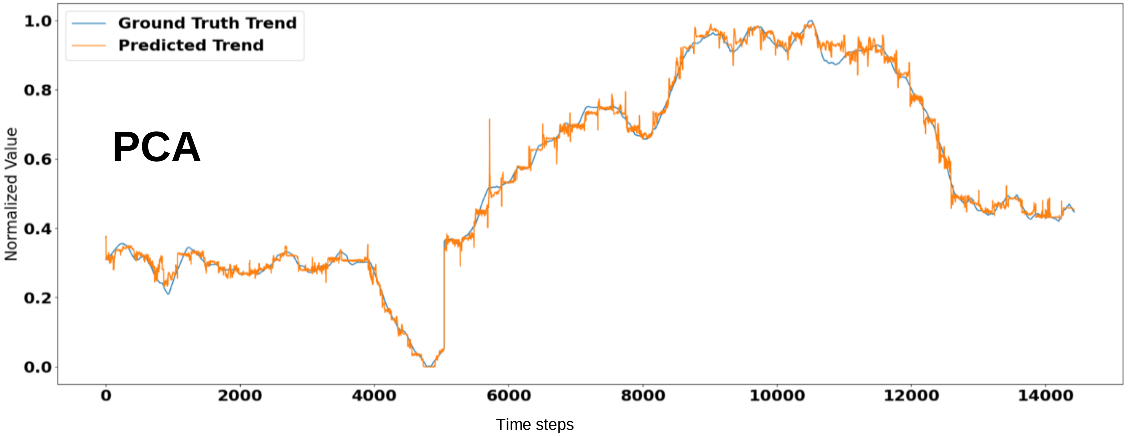
AE – LSTM:

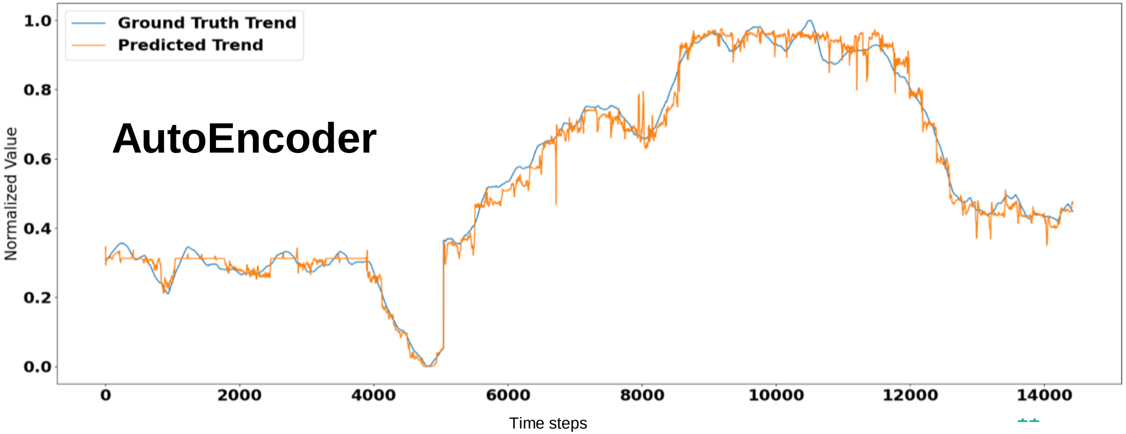


Performance

PCA:
Validation accuracy 4%
AutoEncoder:
Validation accuracy 5 %

Heuristic matrices update every ~7 hours



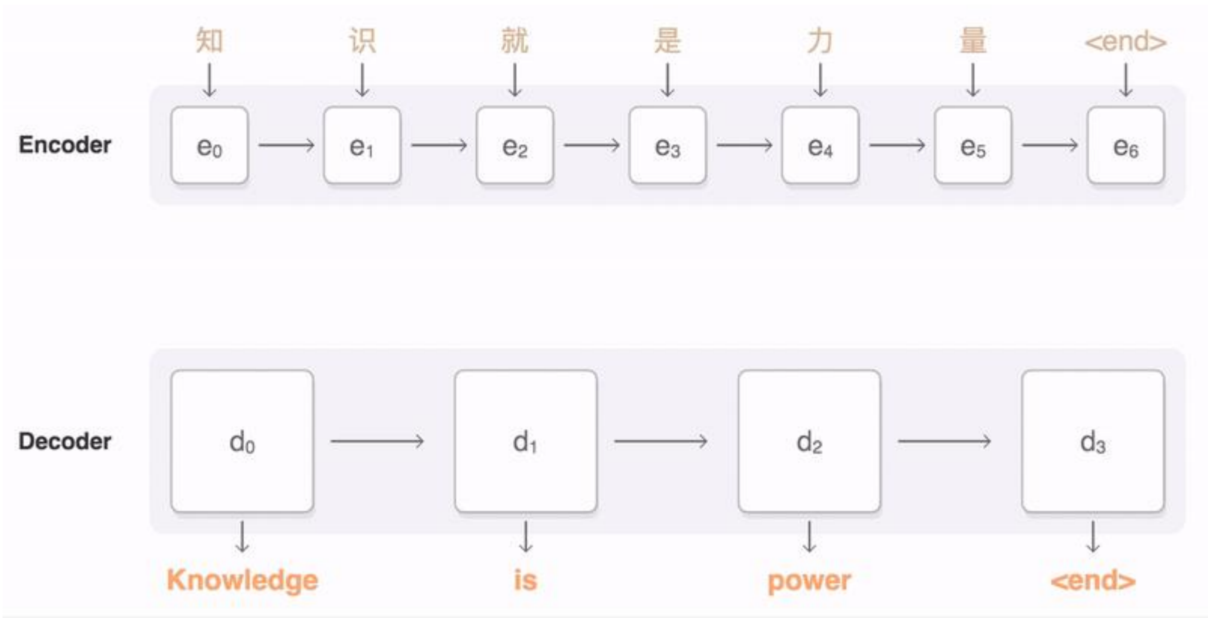


CERN openlab

Forecasting throughput

https://google.github.io/seq2seq/

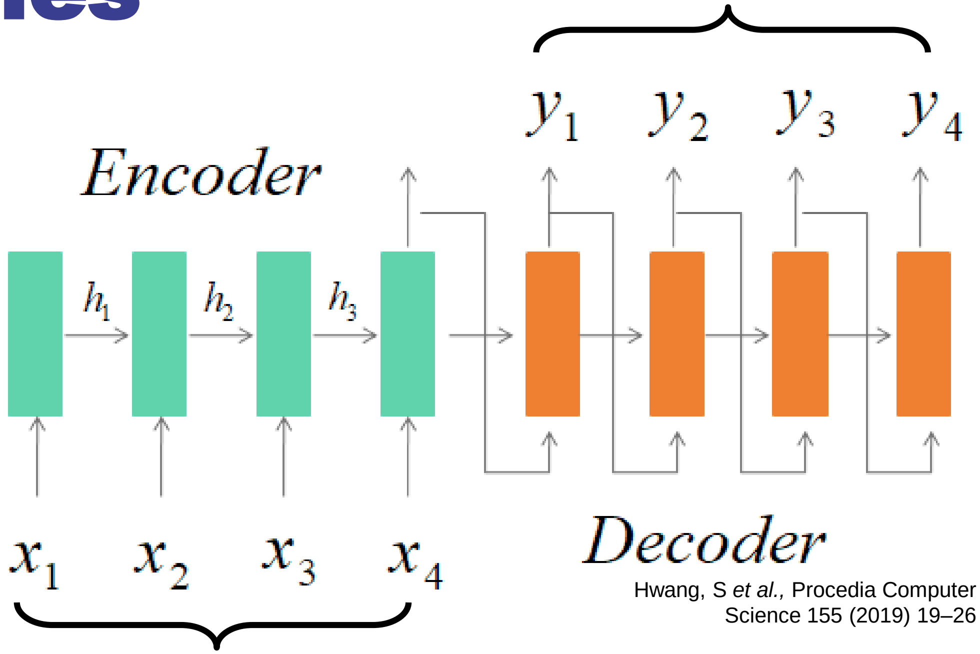
- Build model inspired by Seq2Seq
- Initially introduced for NLP applications
- Drop heuristic matrix information
- Use query readsize values and past throughput



CERN openlab

Seq2Seq time series

- Bin queries RS values in 1 sec intervals per time step
  - 120 RS values in input X_i:
  - $X_n = \{RS_{i,t}, \text{throughput}_t\}$
- Predict subsequent throughput
  - $Y_n = \text{throughput}_{t+n}$
- 2 hyperparameters: input & output sequence lengths



Hwang, S *et al.*, Procedia Computer Science 155 (2019) 19–26

Results

Grid search to optimise input/output size

Best results achieved using low in/out dimensions

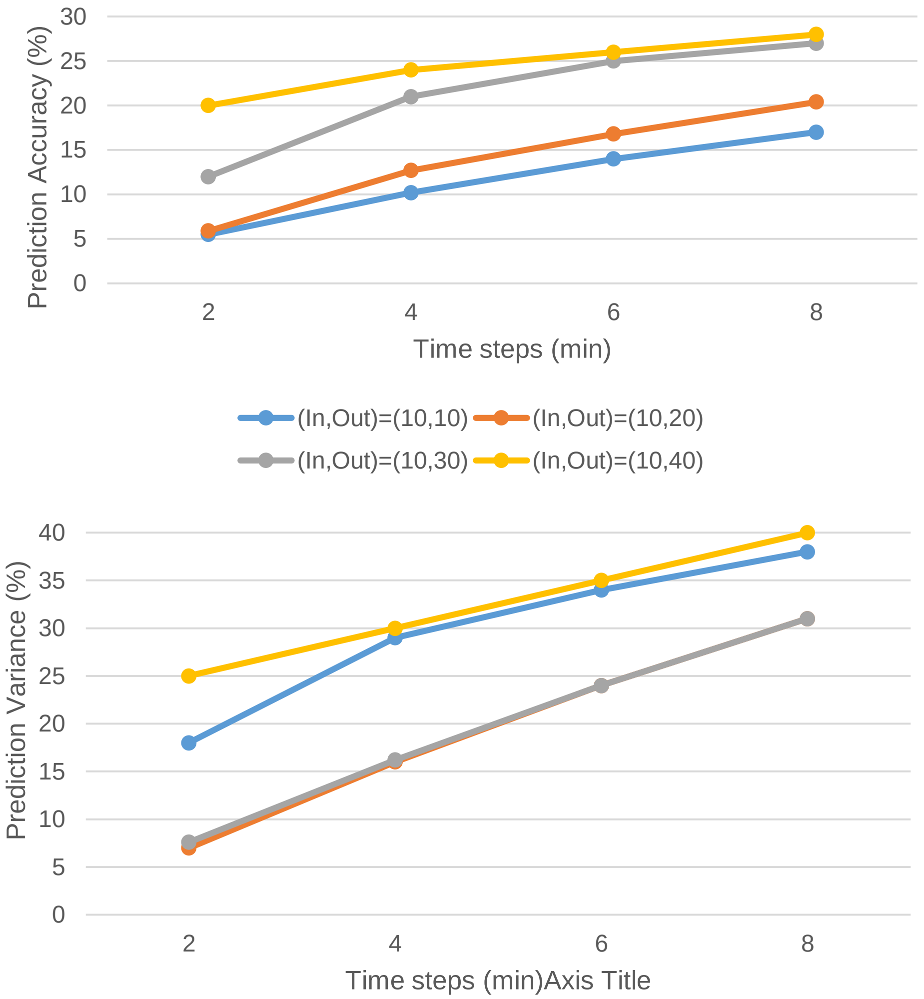
Predict next step (2min) with 5% accuracy

Next step prediction is stable

Preformance degrades when forecasting over longer time spans

8 min forecast → ~15%

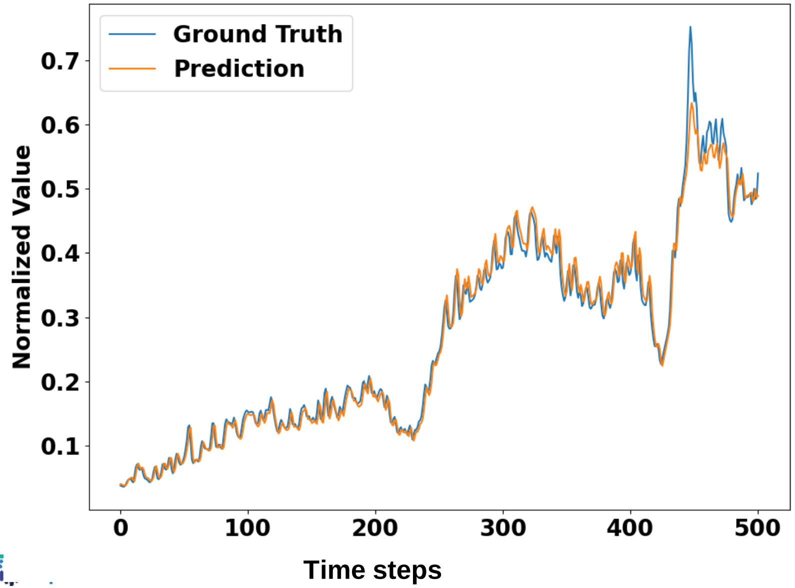
Increasing input size deteriorates the accuracy > 20%

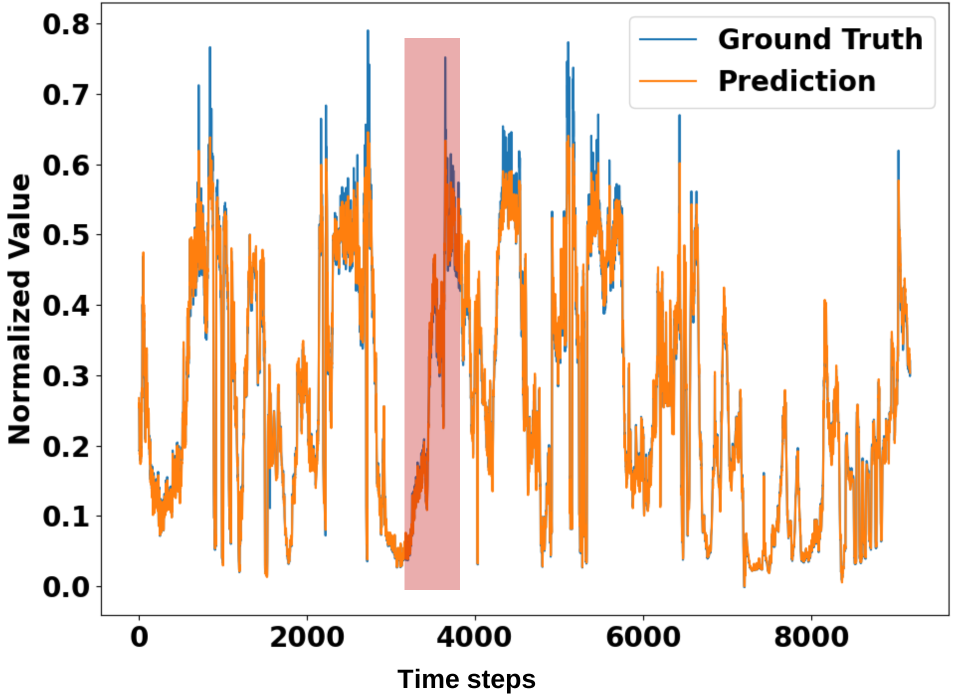


Results

(in,out)=(10,10)

Prediction= t + 2min

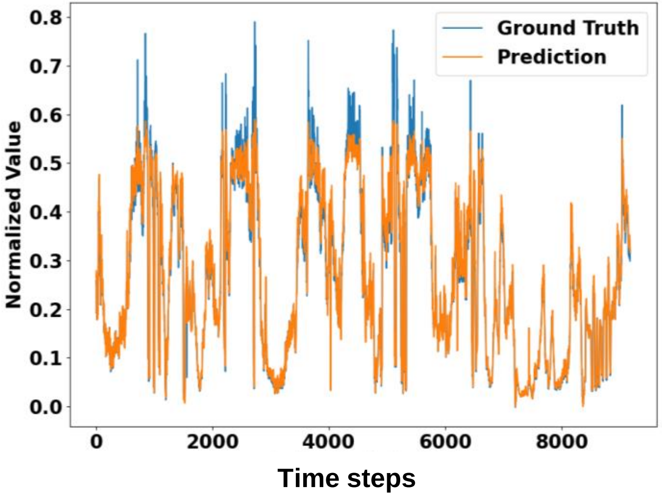




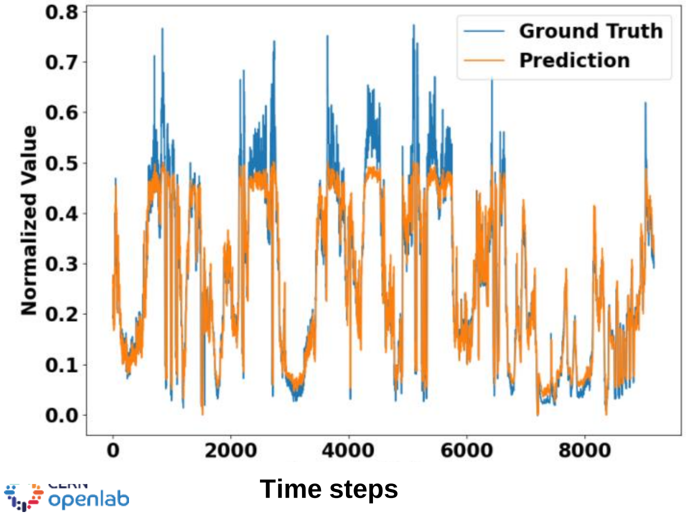
Results

(in,out)=(10,10)

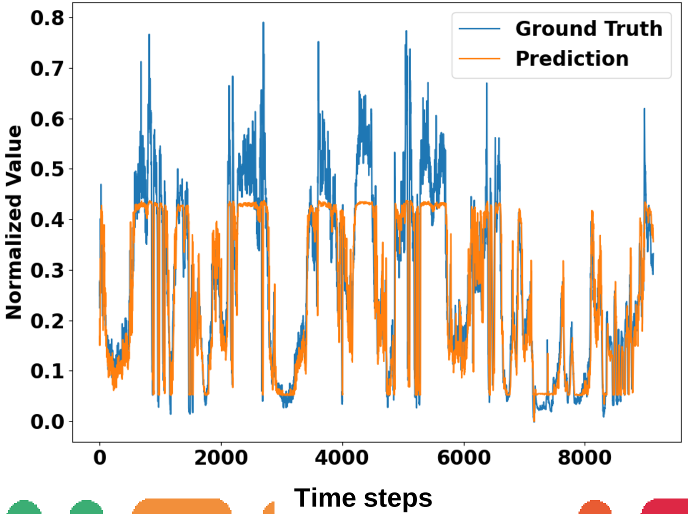
t + 4min



(in,out)=(10,10), **t + 8min**



(in,out)=(30,30), **t + 20min**



Summary

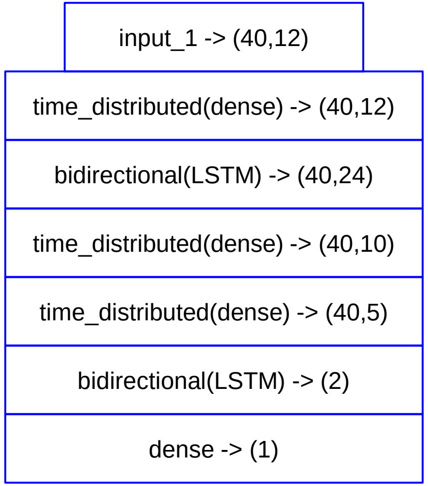
- Investigate Recurrent Neural Networks -based architectures to model and predict I/O throughput
  - **4% accuracy** is a great result for the simulation of such complex dynamic system
  - **Short term (2mins) prediction is good** (5% accuracy)
- Performance decreases in time
  - Current model is very simple: various strategies for improvement
- Accurate prediction of the ALICE jobs I/O behavior can lead to a **reduction of the time to answer client queries**
  - In-memory learned model could be used instead of querying the database.
  - Develop a (quantum) Reinforcement Learning (RL) based optimisation

Thanks!

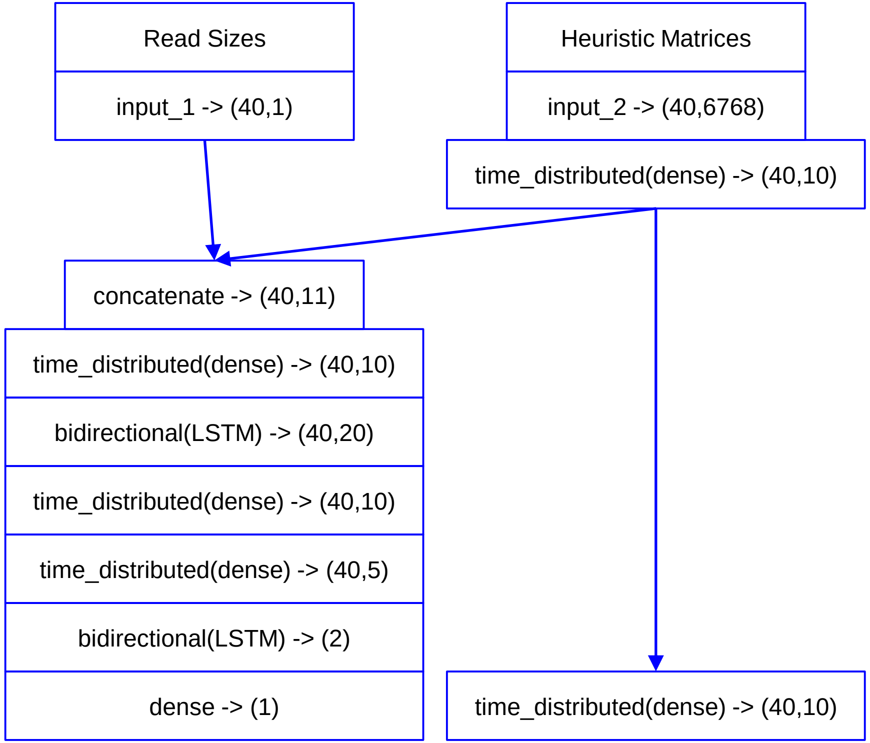
Questions?

PC and ED networks

PCA - LSTM

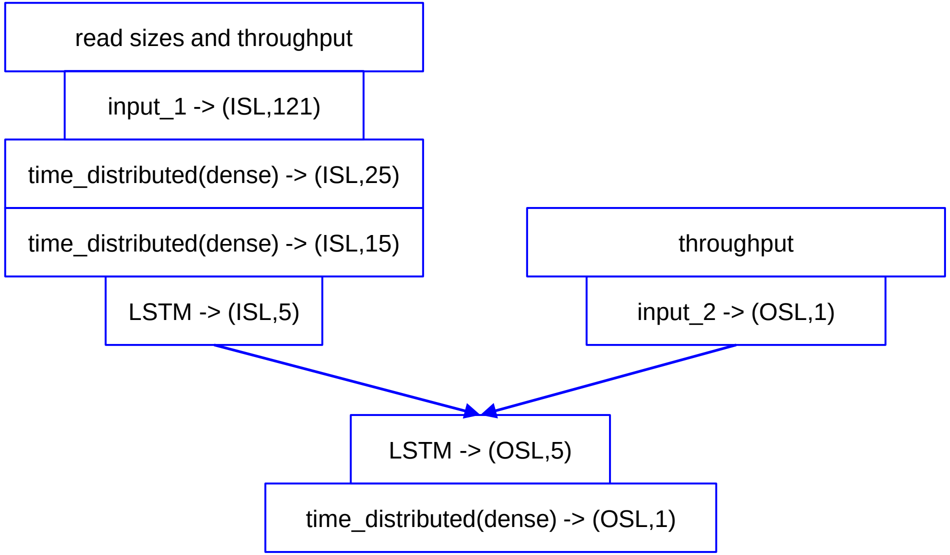


AE - LSTM



Sequence-to-Sequence network

- 2 main customizable hyperparameters
  - input sequence length : ISL
  - output sequence length : OSL
- tested on 2 data sets
  - train and validation data set : TVDS
    - used for the preliminary studies as well
  - control data set: CDS



Validation MAPE

Performance comparison for the first 10 steps between models

- ISL and OSL take values in { 10, 20, 30, 40 } = { 20min, 40min, 60min, 80min}
- 16 models trained

BEST