Spatial-Temporal Explanations for Storage Failure Predictions based on Multivariate Telemetry Sensors

Ioana Giurgiu, Anika Schumann

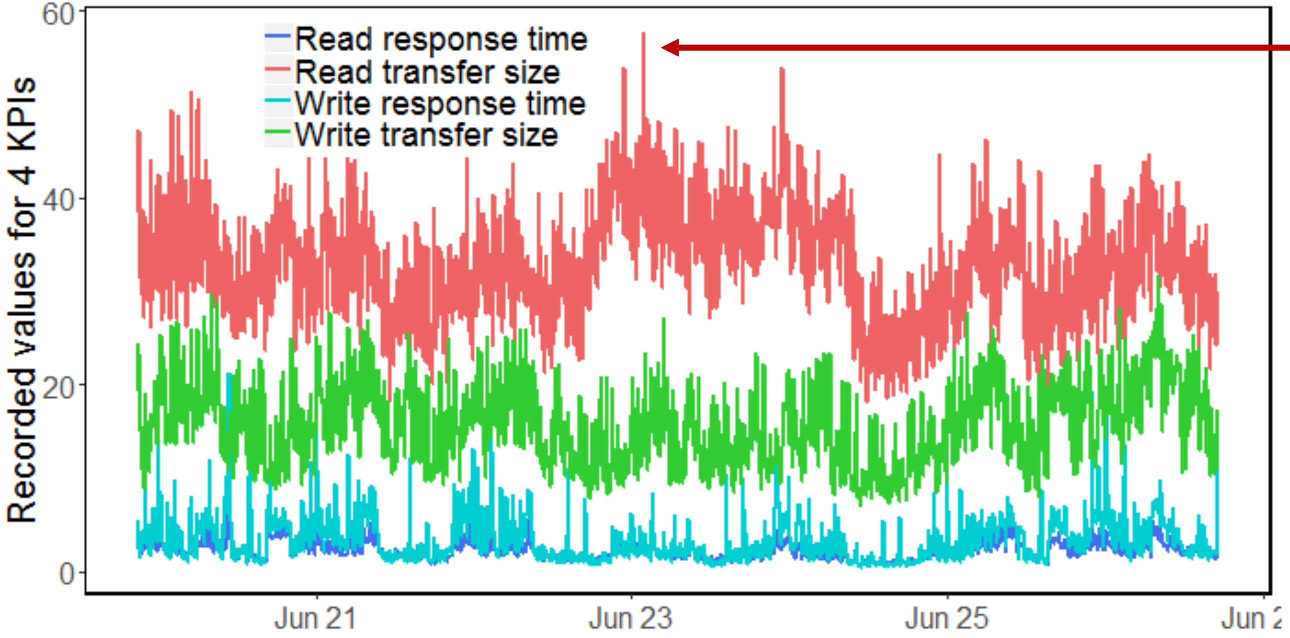
IBM Research – Zurich

Goal

- Explain predicted failures in large-scale real world storage environments based on multivariate telemetry sensors (key performance indicators = KPIs) collected periodically with fine granularity
- Explanations are **spatial-temporal**
- High-level approach:
  - Based on the underlying characteristics of the KPIs, we transform the multivariate time series into **multivariate series of clustered anomalous events of the type KPI_t > threshold**
  - These anomalous events are used in an **LSTM-based network** with **attention** and **temporal progressions** to predict failures 3 days in advance
  - Their **types**, **occurrences** and **frequencies** are used to explain the predicted failures, in both space (which KPIs) and time (when)

Motivation

- Transforming the time series into event series is motivated by the data
  - **KPIs are spiky in nature**, with no increasing or decreasing trends over time



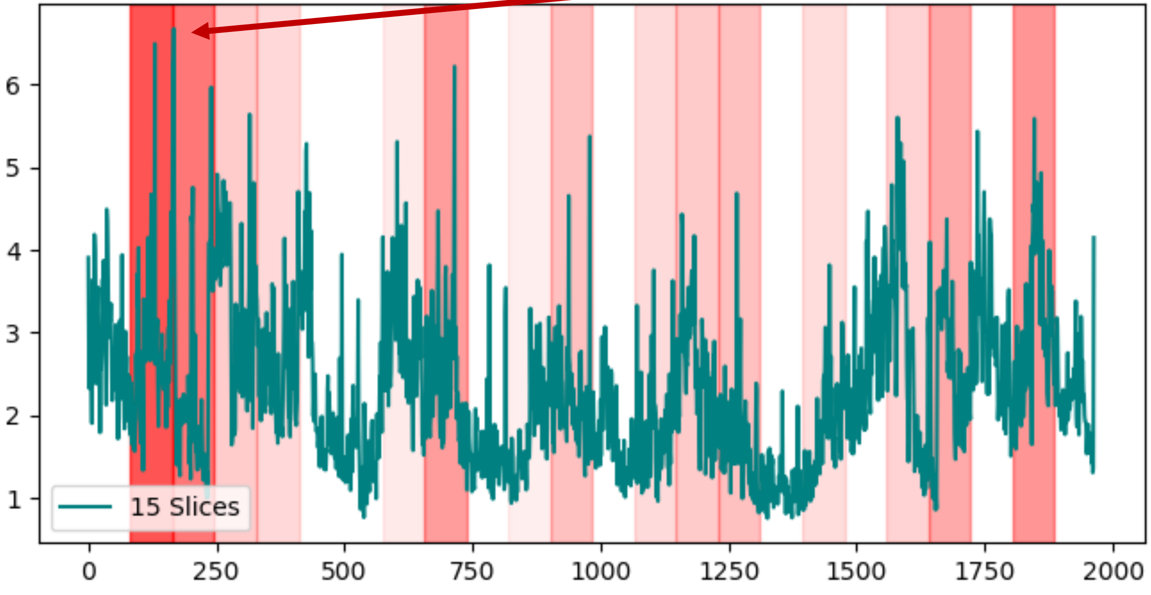
Spikes occasionally exceed pre-defined thresholds

Changepoint detection analysis finds **no significant changepoints** for all KPIs

Motivation (cont.)

- **Model-agnostic explainable approaches do not take the temporal component into consideration**

LIME for time series



Highest contribution is attributed to the earliest slice in the time series (does not reflect a system's behavior)

Quality of explanations highly depends on # slices

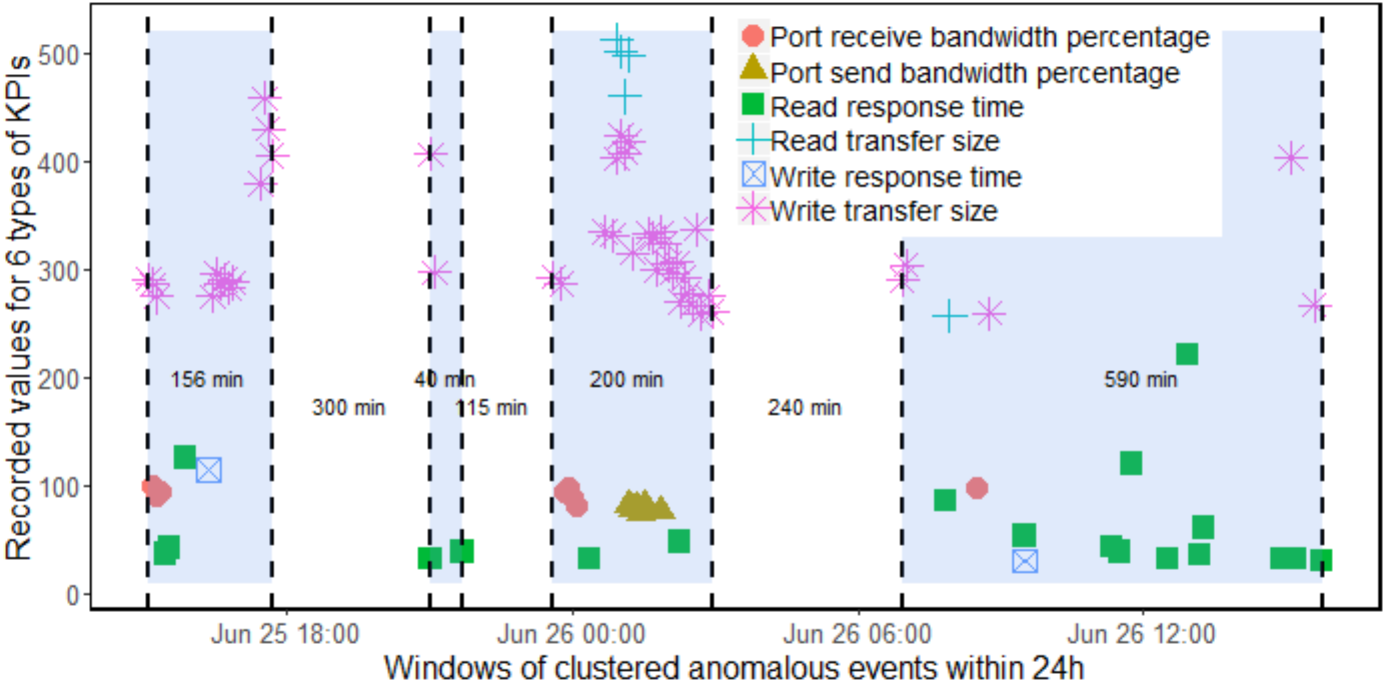
Slices have a fixed length

Fewer slices result in less discrimination in the explanations

More slices result in a vast number of imprecise and misleading explanations

Motivation (cont.)

- **Anomalous events co-occur within well-separated time windows**



Approach

- **Step #1 ➔ Windows of anomalous events** $W_1, \ldots, W_p$ are detected in a time interval [0, t] (observation period) for each storage device in the data set
  - Optimally with Ckmeans.1d.dp

- **Step #2 ➔ Unique anomalous events are embedded** in a continuous vector space as v_e

- Step #3 ➔ For each anomalous event e_n in a window W_r with *N* events, **attention mechanisms aggregate context information** in a context vector:

$$\boxed{cv_{e_n} = \sum_{x=1}^{N} \alpha_{nx} v_{e_x}}$$

Attention value defined as
(Vaswani et al., 2017)

$$\boxed{\alpha_{n1}, \ldots, \alpha_{nN} = \mathrm{softmax}([\frac{q_n k_1}{\sqrt{a}}, \ldots., \frac{q_n k_N}{\sqrt{a}}])}$$

Approach (cont.)

- **Step #4 ➔** For each event, we build a **temporal progression function** that quantifies its impact on the prediction depending on its type and when it occurred:

$$\boxed{I(c_{e_n}, \Delta) = S(\theta_{e_n} - \sigma_{e_n}\Delta) \in [0, 1]}$$

Sigmoid function
(diminishes contributions of events in the distant past)

Initial contribution of e_n

$\Delta = t + T - \zeta W_r$ (time elapsed from Wr to end of prediction window)

Progression of the contribution over time

- **Step #5 ➔** Each window is represented as a **weighted sum of embeddings of its events**:

$$\boxed{w_r = \sum_{n=1}^{N} x_{e_n} I(c_{e_n}, \Delta) c v_{e_n}}$$

How many times event e_n occurred in W_r

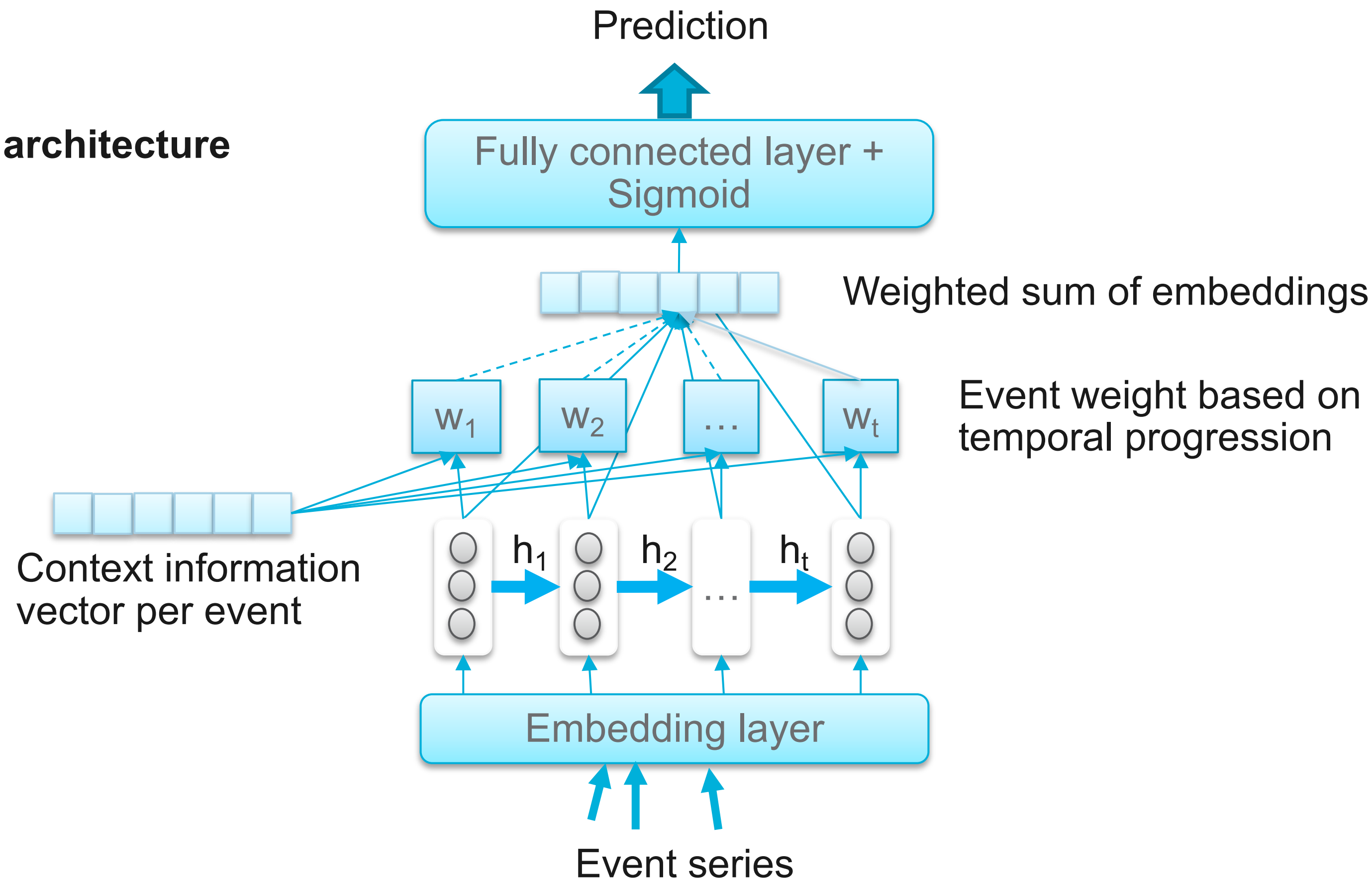
- **Step #6 ➔** The window representations are used in an **LSTM to predict failures**:

$$\boxed{w_r = \sum_{x=1}^{N} \Big(\boxed{\sum_{n=1}^{N} I(c_{e_n}, \Delta) \alpha_{nx}} x_{e_n} v_{e_x} \Big)}$$

Explanations for predictions

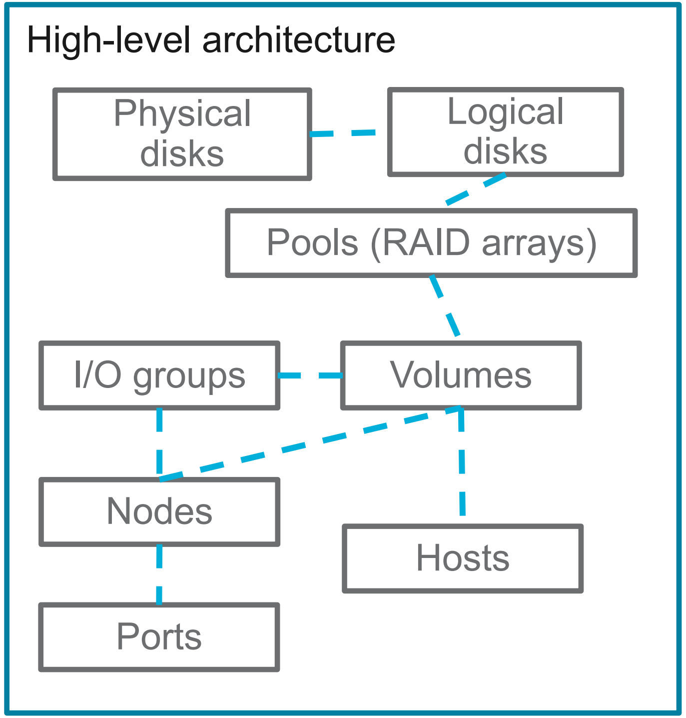
Approach (cont.)

- **High-level architecture**



Data

- **800+ KPIs** collected with 5-min granularity for in 2018 for 130+ storage environments
  - Due to the typical complexity of large-scale storage environments, our dataset consists of over 50 million individual time series
- **266081** anomalous events based on KPI pre-defined rules
- Critical failure incidents used as labels for prediction validation (2% of all incidents)



Settings

- 1:32 ratio between the failure and non-failure classes

- Adam optimizer, batch size = [32,64]
- Initial contribution of event = 1, temporal contribution of event = 0.1
- Dimensionality of event embeddings = 100
- Dimensionality of attention query vectors (q_n) and key vectors (k_n) = 100
- Dimensionality of LSTM hidden state = 100

Results

- **Example #1 ➔ Prediction = Fail** with **0.87** probability

Cluster	Start	Duration	Event	Freq.	Contribution
1	Day 1 22:58	115 min	Read response time Read transfer size Write transfer size	1 5 5	0.00 0.00 0.00
…	…	…	…	…	…
6	Day 5 6:15	120 min	Read response time	2	0.015
7	Day 5 22:55	20 min	Read response time	2	0.02
8	Day 6 22:56	20 min	Read transfer size	1	< 0.01
9	Day 7 23:01	15 min	Read transfer size	2	0.01
10	Day 8 6:02	125 min	Disk utilization	3	0.00
11	Day 8 22:57	20 min	Read transfer size Write transfer size	5 4	0.05 0.16
12	Day 9 23:12	65 min	Read response time	3	0.06
13	Day 11 20:28	205 min	**Write response time**	4	**0.18**
14	Day 13 4:08	35 min	Read response time **Write response time**	4 2	0.1 **0.34**
15	Day 14 22:59	15 min	Read response time **Peak backend write response time** **Write response time**	3 2 3	0.12 **0.8** **0.63**

Results (cont.)

- **Example #2 ➔ Prediction = No fail** with **0.77** probability

Wndw	Start	Event	Frequency	Contribution
1	Day 1 10:07	Disk utilization	1	0
…	….	…	…	…
6	Day 11 18:22	Read transfer size	2	0.05
7	Day 13 2:47	Read response time Disk utilization	2 3	0.04 0.02

Results (cont.)

- **Example #3 ➔ Prediction = No fail** with **0.69** probability

Wndw	Start	Event	Frequency	Contribution
1	Day 2 15:17	Peak backend write response time Read response time	2 3	0.05 0
2	Day 5 12:02	Peak backend write response time	2	0.06

One of the driving metrics shows anomalous events early and not in combination with other driving metrics

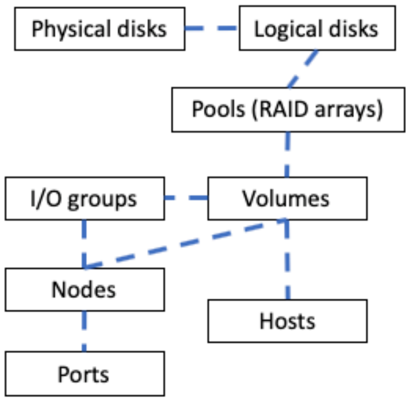
Interactions between metrics and their temporal progression is considered when building the explanations

2-step snapshot

Step 1: Predictions list

Storage ID	Customer	Failure Risk
abc123	XYZ	Critical
jdhf3874	XYZ	Moderate
dsgh343	ABC	Critical
djsj87	XYZ	Critical
jdjhf875	ABC	Moderate
65356	XYZ	Moderate
854mfm	XYZ	Low

High-level architecture



Step 2: Explanations per prediction

abc123 [XYZ]

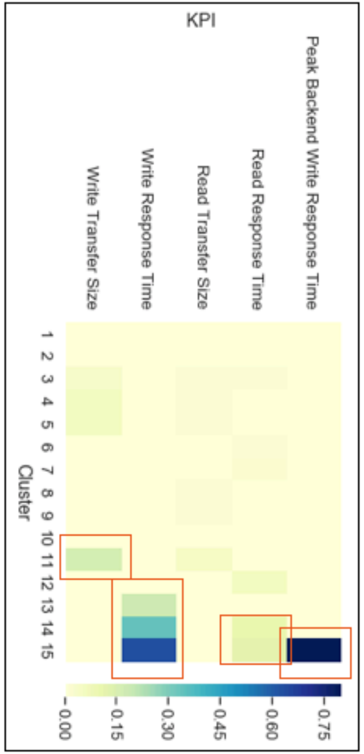
Probability: 87%

0-50% | 50-85% | 85-100%

Events and their contribution to explaining the failure prediction [Table]

Cluster	Start	Duration	Event	Freq.	Contribution
1	Day 1 22:58	115 min	Read response time Read transfer size Write transfer size	1 5 5	0.00 0.00 0.00
...	...	...	...	...	...
6	Day 5 6:15	120 min	Read response time	2	0.015
7	Day 5 22:55	20 min	Read response time	2	0.02
8	Day 6 22:56	20 min	Read transfer size	1	< 0.01
9	Day 7 23:01	15 min	Read transfer size	2	0.01
10	Day 8 6:02	125 min	Disk utilization	3	0.00
11	Day 8 22:57	20 min	Read transfer size Write transfer size	5 4	0.05 0.16
12	Day 9 23:12	65 min	Read response time	3	0.06
13	Day 11 20:28	205 min	**Write response time**	4	**0.18**
14	Day 13 4:08	35 min	Read response time **Write response time**	4 2	0.1 **0.34**
15	Day 14 22:59	15 min	Read response time **Peak backend write response time** **Write response time**	3 2 3	0.12 **0.8** **0.63**

[Heatmap]



Summary

- **Goal**: **Spatial-Temporal explanations** for predicted failures in storage environments on **multivariate time series data**
  - Agnostic explainable models do not take the temporal component into consideration
  - Exploit the spiky nature of the data with anomalous event series extracted from the original time series
- **LSTM + attention + temporal progressions** to predict and explain how each event depending on its type, frequency and occurrence contributed to the failure event
- Explanations are **easy to read** and **understand**
- For time series, **explanations need to be validated by an SME**
  - Essential to present enough explanations to an expert to enable trust in the model
  - … but without providing an overwhelming volume of explanations

Thank you! Questions?

igi@zurich.ibm.com

https://www.zurich.ibm.com/predictivemaintenance/