Information Modelling and Knowledge Bases XXXVI
Y. Kiyoki et al. (Eds.)

doi:10.3233/FAIA241577

Data-Driven Fault Localization in Cyber-Physical Systems Using Dependency Graphs and Anomaly Detection

Arne GRÜNHAGEN [a,b,c,1], Annika EICHLER [b,c] Marina TROPMANN-FRICK [a] and Görschwin FEY [b]

[a] *Hamburg University of Applied Sciences, HAW, Germany*
[b] *Hamburg University of Technology, TUHH, Germany*
[c] *Deutsches Elektronen-Synchrotron DESY, Germany*

Abstract. The early and automatic detection of faulty behavior is essential for maintaining the reliability of a cyber-physical system. In this paper we describe a fault localization approach for such a highly complex distributed system, the optical synchronization system of the European X-ray free-electron laser. Using a dependency graph, we model the relationships between the components and the influences of environmental effects. After we first resolve linear long-term dependencies between dependent components with a correlation analysis, we then use an unsupervised fault detection pipeline consisting of statistical feature extraction and unsupervised anomaly detection to accurately identify anomalies and localize their origins in the system.

Keywords. Data Mining, Fault Analysis, Dependency Graph, Cyber-Physical System

1. Introduction

The optical synchronization system of the European X-ray Free-Electron Laser (EuXFEL) [1] (see Figure 1) consists of two redundant main laser oscillators (MLO) both emitting a laser pulse train with a pulse repetition rate of 216.667 MHz and a pulse duration of 200 fs. The phases of the MLOs are actively stabilized with respect to the 1.3 GHz RF Main Oscillator (MO) using a Proportional Integral (PI) controller in a phase-locked loop with a loop bandwidth in the order of 1 kHz to 10 kHz [2].

The pulse train from the MLO is split and transmitted to various fiber link stabilizing units (LSU) with active length stabilization. Optical fibers are employed to establish connections between the LSUs and the respective end stations in the accelerator, such as laser synchronization setups, the RF re-synchronization units, and bunch arrival time monitors (BAM). Furthermore, a sub-distribution system is set up in the experimental

[1] We acknowledge the support by DASHH (Data Science in Hamburg - HELMHOLTZ Graduate School for the Structure of Matter) with the Grant-No. HIDSS-0002.

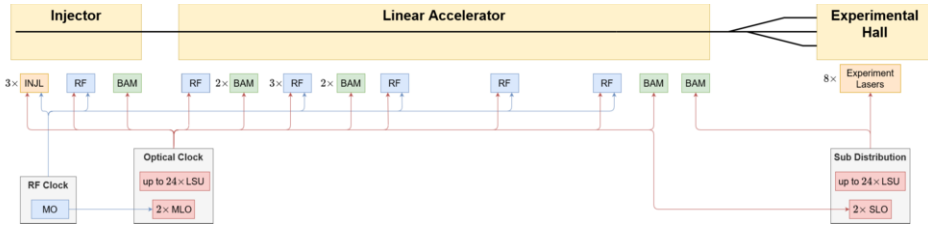


Figure 1. Schematic Overview of the Optical Synchronization System of the European XFEL [3]

hall 3.4 km away, in which another two laser oscillators are actively phase-locked to the MLO. This arrangement allows the sub-distribution laser oscillators (SLO) to replicate the MLO signal and to distribute the optical synchronization signals to additional end stations. This large setup ensures synchronization on a femtosecond level between the accelerator components and the various experimental lasers, enabling advanced research at the EuXFEL.

Ensuring the continuous availability and optimal performance of such cyber-physical systems poses significant challenges. Experiments can only be conducted effectively when the synchronization system operates reliably at peak performance levels. Manual inspection of the numerous system components is impractical due to their scale, number, and distributed nature, necessitating an automated approach for comprehensive monitoring. Many of the distributed subcomponents of the optical synchronization system are interdependent. For example, if a phase shift occurs at the MLO, this phase shift is passed on to the synchronized end stations via the respective links. The optical synchronization system is therefore regarded as a distributed system of interdependent components. Furthermore, the individual components are also dependent on environmental influences.

The aim of this work is to create a fault localization system for the components of the optical synchronization system, taking into account the interdependencies between the various system components and the influences from environmental changes. The fault localization will be used to characterize how components react to the behaviour of other components or environmental influences.

The paper is structured as follows: in Section 2 we describe work that also deals with fault localization and anomaly detection in complex cyber-physical systems, Section 3 gives a detailed overview of the analyzed data and the used methods. The results are presented in Section 4. Finally, we summarize and evaluate the presented methods and results in Section 5 and give an outlook for future work.

2. Related Work

The authors of [4,5] give a detailed survey about fault diagnosis for industrial systems. They give an overview of data-driven methods, model-based methods, signal processing methods and hybrid methods. They also relate the methods described to distributed system structures that are similiar to the distributed nature of our system. It is also relevant for our work that the fault diagnosis methods are evaluated on data in the time domain and frequency domain.

The authors of [6] deal with fault detection and localization for large scale power systems. By combining change point detection and a dependency graph, the complex system is visualized and then examined for faults. Their main focus is on decentralizing the computation of the fault analysis to the respective components. In [7] the authors describe a fault detection system based on an automatic feature extraction using a Convolutional Neural Network (CNN) autoencoder and fault detection based on Bayesian change point detection. The proposed pipeline was evaluated on different data sets of distributed systems. The authors of [8] also deal with the problem that dependencies in distributed complex systems make it difficult to isolate anomalies. For this purpose, the anomaly attributed on one component is compared with possible anomalies of neighboring components. In our work, we use not only the locally neighboring components but also the logical relationships between the components that are far apart. In [9] the authors present a Gaussian model based fault diagnosis for the low level radio frequency subsystem of the EuXFEL. This is interesting for us as we have to deal with the same facility, similar dependencies, and external conditions. However, the subsystem and the data are different in detail.

In conclusion, there is previous work dealing with fault diagnosis of distributed systems and each of the presented papers addresses the issue of dependencies between components or the complexity of the systems. In our work, we will also represent the dependencies of the underlying distributed cyber-phasical system using a dependency graph. However, our focus is on determining the dependencies of the components' behavior and not on decentralized computation.

3. Data and Method

This section describes both the data used and the methodological part of this work, with which a fault localization is carried out.

3.1. Optical Synchronization System as a Dependency Graph

The optical synchronization is a network of interdependent components, as illustrated by the dependency graph $G = (V, E)$ as depicted in Figure 2.

- V represents a set of vertices within the graph. Each vertex $v_i \in V$ represents an individual component of the system or a specific environmental component.
- $E \subseteq V \times V$ represents a set of directed edges between vertices. Each directed edge $(v_i, v_j) \in E$ represents a dependency relationship between vertices v_i and v_j, where the behavior of v_i depends logically on v_j.
- The behavior of each component v_i is characterized by a set of data channels $D_i = \{d_{i1}, ..., d_{in}\}$, where d_{ij} represents the j-th data channel belonging to the i-th component v_i.
- Each data channel of the component v_i has the same set of influencing components and therefore also the same set of influencing data channels that is defined as $I_i = \cup_{(v_i,v_j)\in E} D_j$.

The MLO is phase-locked to the signal of the MO. The SLO is phase-locked to the outgoing signal of the MLO. This results in direct dependencies from the MLO to the

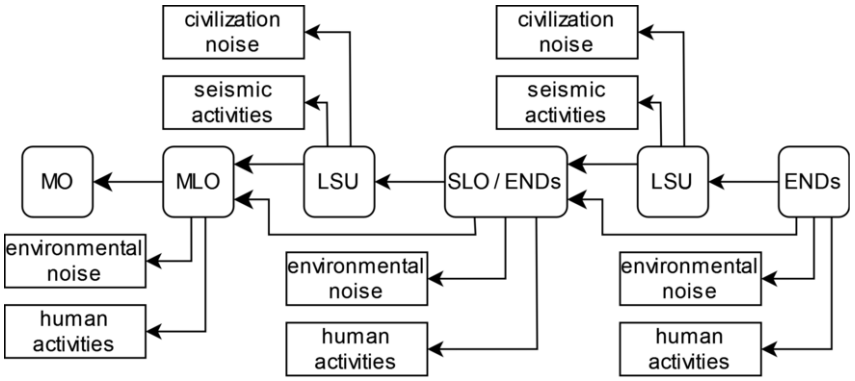


Figure 2. Dependency Graph of the Optical Synchronization System

MO and from the SLO to the MLO, as well as an indirect dependency from the SLO to the MO. The end stations ENDs (REFM-OPT, experiment laser, BAM) are synchronized either by the MLO or by the SLO. This also results in direct and indirect dependencies between the components. The synchronization signal is transmitted via optical fibers between the MLO and SLO as well as between the laser oscillators and the respective ENDs. These optical fibers are length-stabilized in the LSUs via an optical cross-correlator and active control by PI controllers [10]. Phase noise, which is generated by the laser oscillator, also has an influence on the length measurement of the LSU. Therefore, there are also direct dependencies between the LSUs and the respective laser oscillators.

In addition to the dependencies on other system components, the laser oscillators are also dependent on environmental variables. Laser oscillators can be strongly influenced by events in the environment like human activities or environmental noise. For example, in [11] it is shown how a laser oscillator reacts to acoustic disturbances. Furthermore, changes in temperature, humidity and air pressure can have an influence on the performance of the laser oscillator. For these reasons, the laser oscillator setup is placed in a temperature and humidity controlled environment. However, this does not protect the system from changes in air pressure.

Due to seismic activities such as earthquakes or underground ocean waves, the distance between two synchronized components changes. These external changes in length are compensated for by the active length stabilization of the LSU. The EuXFEL tunnel is located below a residential area. Therefore, vibrations caused by civilization can also have an influence on the tunnel and thus on the LSU behavior.

3.2. Available Data

Components of the Optical Synchronization System

A stable phase of the laser oscillators is crucial for operating the system. The phase of the respective outgoing signals is stabilized in a phase-locked loop (PLL) using a proportional-integral controller. Therefore, the controller input and output signals are utilized to describe the behavior of the laser oscillators. The PLL consists of the following components:

1. **Phase detection:** The phase detector compares the phase difference between the reference input signal ($\theta_{\text{ref}}(t)$) and the feedback signal from the laser oscillator ($\theta_{\text{feedback}}(t)$) leading to the phase detector output:

$$\Phi_{\text{error}}(t) = \theta_{\text{ref}}(t) - \theta_{\text{feedback}}(t) \tag{1}$$

2. **Loop filter:** The PI controller consists of a proportional gain (K_p) and an integral gain (K_i).

$$V_p(t) = K_p \times \Phi_{\text{error}}(t) \tag{2}$$

$$V_i(t) = K_i \times \int \Phi_{\text{error}}(t)\, dt \tag{3}$$

3. **Laser oscillator:** The laser oscillator reacts to the controller generated output signal V_c combining the proportional and the integrator parts:

$$V_c(t) = V_p(t) + V_i(t) \tag{4}$$

In this paper we use the controller input $\Phi_{\text{error}}(t)$ and the controller output $V_c(t)$. Similar to the approach taken for the laser oscillators, we also employ controller input and controller output signals to characterize the LSUs. This methodology ensures a comprehensive analysis of both laser oscillators and LSUs, emphasizing the significance of stable phase conditions in assessing their overall performance.

Environmental Influences

In addition to the data sources of the system components, we use sensor data describing the temperature, relative humidity and air pressure in the laboratory of the laser oscillators and the LSUs. The optical fibers, which connect components over a distance of up to 3.4 km, are guided through the accelerator tunnel. Due to seismic activities, there are small deformations of the tunnel and thus changes in the length of the synchronization path. These environmental influences have a negative impact on the performance of the optical synchronization system because their effects are mostly below the control bandwidth of the PI controller. For this reason, we use the following databases that would indicate seismic activities near the EuXFEL tunnel:

- U.S. Geological Survey earthquake database [12]
- traffic data above the accelerator tunnel [13]
- weather data describing the sea level in the North Sea [14]

Figure 3 shows where the number of cars passing the marked intersection in the direction of EuXFEL are counted and Figure 4 shows the positions of the sea buoys measuring the sea state.

3.3. Fault Localization

The logical dependencies between the components indicate that certain behavior in one component has an influence on data channels in dependent components. These dependencies are divided into two categories:

- The long-term trend of a data channel is determined by other component's data channels at all times.

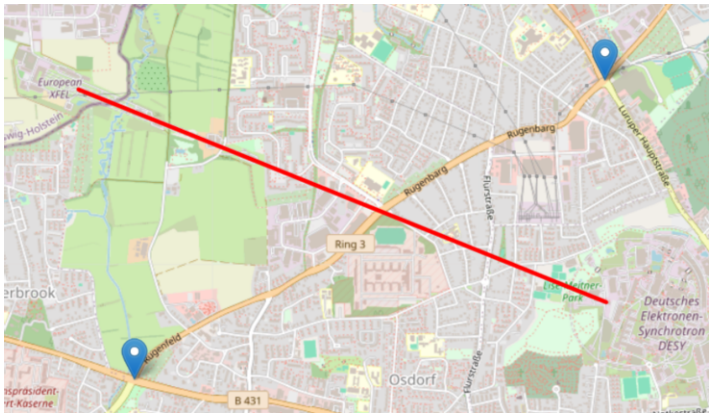


Figure 3. Positions of the infrared cameras (blue) for counting cars in relation to the European XFEL (red) [15]

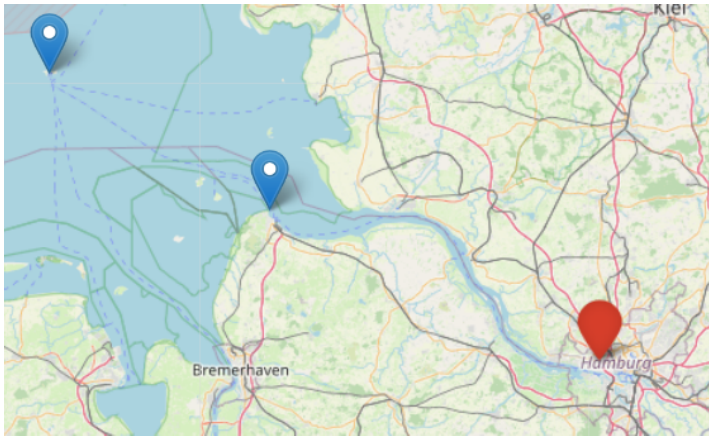


Figure 4. Positions of the weather stations (blue) in relation to the European XFEL (red) [15]

- Spontaneous anomalous effects in a data channel of a component, which may indicate faulty behavior, also trigger anomalous behavior in data channels of a dependent component.

The dependencies are analyzed in two steps, as shown in Figure 5. First, the long-term behavior of the signals is examined to see whether the trend of a data channel is reflected in a dependent data channel. If this is the case, this trend is determined and subtracted. In the second step, the cleaned data channels are examined for anomalies and non-linear correlations and then the origin of the anomaly is localized with the help of the dependency graph.

Elimination of Long-Term Linear Correlations

In this section, the process of eliminating long-term trends is described and illustrated in Figure 6. The result of this process is that linear long-term effects, which correlate with the trend of influencing signals, are removed. This process is carried out for data channels belonging to components of the optical synchronization system. The set of data

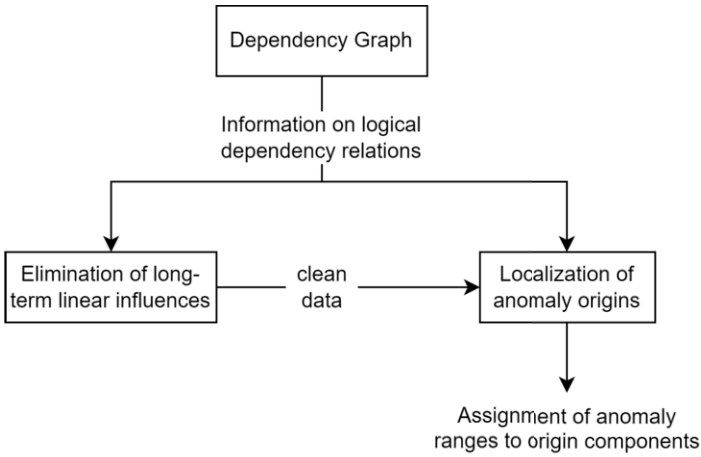


Figure 5. Methodology of data driven fault localization

channels describing the k-th system component $D_k = \{d_{k1}, ...d_{kn}\}$ is transformed into a correlation-free version $D'_k = \{d'_{k1}, ...d'_{kn}\}$. An important assumption for the described trend elimination is that the trend influences of different data channels are independent of each other and in combination do not cause a different behavior in the target data channel. The following is a step-by-step description of how a target signal is cleaned of all the influencing signal trends.

1. **Trend Analysis:** First, the trend of the target signal d_{tk}, which is the k-th data channel of component t, and the set of all influencing signals I_t are determined. To do this, each signal is split into equally sized windows and the median of each window is calculated. This process, also known as median smoothing, isolates the low-frequency trend of the target signal from general noise. Positive and negative peaks that have not been removed by median smoothing are then removed by a peak detection. The same process is carried out for the influencing signals.
2. **Correlation Analysis:** Pearson correlations [16] are calculated between the trend of the target signal and the trends of all influencing signals.
   The Pearson correlation coefficient p is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. The Pearson correlation coefficient between two variables X and Y with n paired data points is given by:

$$p_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{5}$$

   Where X_i and Y_i are individual data points for variables X and Y, respectively. $\bar{X}$ and $\bar{Y}$ are the means of variables X and Y, respectively. The Pearson correlation coefficient can range from -1 to 1:

   - A value of 1 indicates a perfect positive linear relationship, implying that as one variable increases, the other variable also increases proportionally.
   - A value of -1 indicates a perfect negative linear relationship, suggesting that as one variable increases, the other variable decreases proportionally.

- A value of 0 indicates no linear relationship between the variables.

The influencing signal with the highest Pearson correlation to the target signal is used for cleaning the target signal if its correlation exceeds a predefined threshold. If no correlations of the influencing signals exceed the threshold value, this means that the influencing signals do not have a linear relationship to the target signal.

3. **Scaling:** Scaling is performed both for the target signal and for the influencing signal. This step is crucial for the following signal alignment, which involves distance measurements between the signals.
4. **Signal Alignment:** To remove the trend of the influencing signals from the target signal, it is important that the signals are perfectly aligned. This alignment is carried out in two steps. First, the trend of the influencing signal is shifted over the trend of the target signal. This determines at which shift the signals have the highest Pearson correlation to each other. After this alignment, the signals are fine aligned at the local level using dynamic time warping [17] and a Sakoe-Chuba band [18] that limits the maximum shift of each individual datapoint to 10 min.
5. **Linear Regression:** The aligned signals are reverse scaled and used to create a linear regression model. This model transforms the trend of the influencing signal to the trend of the target signal.
6. **Trend Removal:** Utilizing the linear regression model, the correlation component is calculated, and a correlation-free version of the target component's signal is derived.
7. **Iterative refinement:** The pipeline iterates using the correlation-free version of the target's signal, excluding previously used signals. After this iterative refinement, there are no longer any high correlations between the trends of the system data channels and the trends of the respective influencing data channels.

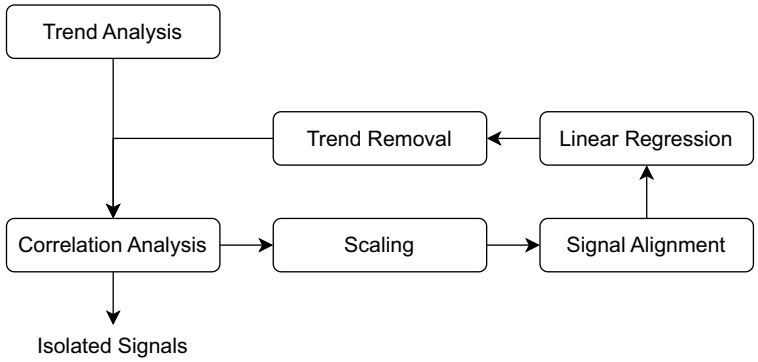
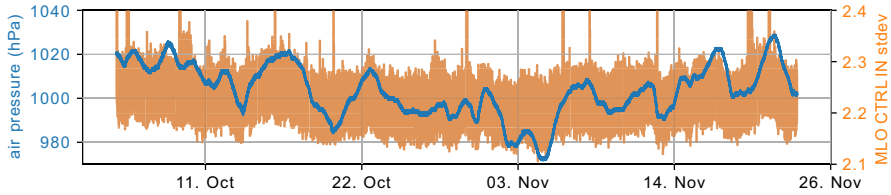
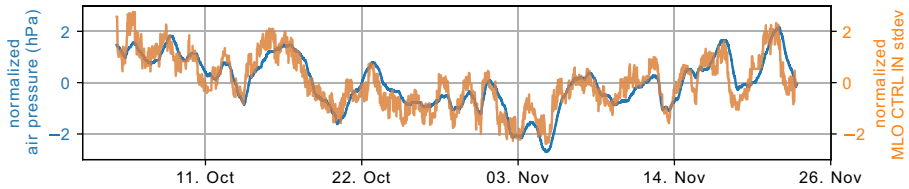


Figure 6. Process of eliminating of long-term linear correlations

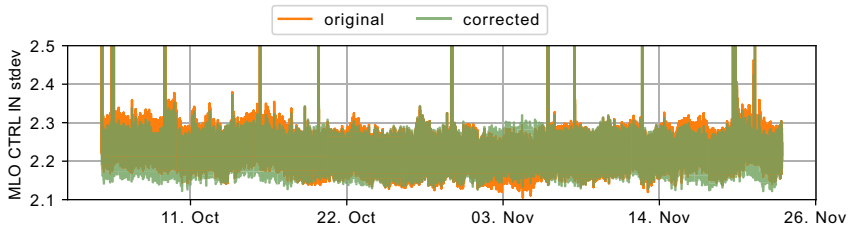
An example of the correlation elimination described is shown in Figure 7. The phase noise of the MLO, which serves as a controller input, is cleaned from the influence of the air pressure.



(a) Measured laboratory air pressure and calculated standard deviation of the MLO controller input (MLO CTRL IN stdev)



(b) Aligned and scaled trend data



(c) Removed air pressure trend from controller iput standard deviation

Figure 7. Example of the eliminating long-term linear dependencies

3.4. Unsupervised Anomaly Localization

Data-based anomaly localization describes the process of identifying data points that deviate significantly from typical patterns within a data set and subsequently determining the origin of the atypical behavior. A complex system such as the optical synchronization system has many components, some of which are dependent on each other or on external environmental influences. This means that anomalous behavior may propagate from one component to dependent components. The following pipeline is used to determine the origin of anomalous behavior at a specific target component. After this step each data channel, including system data channels and environmental data channels, gets assigned a set of n anomalies $A_i = \{a_1, ..., a_n\}$, where an anomaly is characterized by a time range and an origin, which are determined in this step. This process is based on the assumption that there is a causal relationship between an anomaly in both the target data channel and the influencing data channel. We use a combination of statistical methods and unsupervised outlier detection to identify and describe anomalous behavior.

1. **Data Input:** The input to the pipeline is the target data channel d'_{ik}, which has been cleaned from long-term linear correlations, and the set of all influencing signals I_i .

2. **Data segmentation:** The input data is split into non-overlapping windows of a predefined size.
3. **Feature extraction:** For each window, a set of diverse statistical features is computed to characterize the data's behavior. These features include mean, standard deviation, skewness, kurtosis, and providing a comprehensive representation of the window's content.
4. **Anomaly detection:** An algorithm for detecting outliers is used to assess whether the data points of the windows exhibit anomalous behavior. This algorithm generates a label for each window to indicate whether it is anomalous or not. We used the anomaly detection algorithm Local Outlier Factor [19]. The identical procedure is repeated for all influencing signals of the target signal. If the outlier ranges are directly adjacent, these ranges are merged. The result of this step is that a set of anomalies is assigned to each data channel.
5. **Localization of anomaly origin:** The identified outlier ranges in the target data channel d'_{ik} are compared with those found in the data channels of the influencing components. For the data channels whose component v_i has no influencing signals, i.e. $I_i = \emptyset$, the origin of all detected anomalies is set to the respective component v_i. If an outlier range in the target data channel matches an outlier range in a data channel of an influencing component v_j, this indicates that the outlier has propagated from v_j to v_i. The origin of the anomaly of the influencing component is adopted. If, on the other hand, no matching outlier ranges are found, it is assumed that the outlier originates either from the component itself or from an unknown source, it is called independent outlier and again the origin is set to the component v_i itself.

4. Results

This section describes which dependencies and influences were identified with the help of the methods described in Section 3. For each component of the optical system we evaluate, to what extent it depends on other components or external environmental influences and which anomalies are introduced into the respective component by other components or environmental influences. Furthermore, the anomalies that are not brought into a component from the outside are also analyzed. These anomalies are assumed to originate either from the analyzed component itself or from data sources that are not yet available.

As described in Section 1, the optical synchronization system consists of an MLO that repeats the reference of the MO, the SLO that repeats the reference of the MLO and other end stations that are either synchronized by the MLO and a link or via the SLO and a link. The optical synchronization system has this pattern consisting of synchronizing laser oscillator, synchronized component and connecting link unit several times. In the following, the results between the MLO and the SLO as well as the connecting link are shown in detail. The analyzed data set contains the respective data channels in the months of October and November. Instead of the exact controller input and output signals, which would each have a resolution of 0.3 MHz, the mean value and the standard deviation of 0.1 s windows are used. In addition, low frequency phase drift from 0 Hz to 5 Hz were calculated from the respective mean data channels of the controller outputs and inputs

using 10 min Von-Hann windows, Fourier transformations and numerical integration. The full list of data channels used is depicted in Table 1.

Table 1. Available data channels of system components

System Component		Data Channel
MLO controller	input	mean (CTRL IN mean)
		standard deviation (CTRL IN std)
		phase drift (CTRL IN phase drift)
	output	mean (CTRL OUT mean)
		standard deviation (CTRL OUT std)
		phase drift (CTRL OUT phase drift)
SLO controller	input	mean (CTRL IN mean)
		standard deviation (CTRL IN std)
		phase drift (CTRL IN phase drift)
	output	mean (CTRL OUT mean)
		standard deviation (CTRL OUT std)
		phase drift (CTRL OUT phase drift)
LSU controller	input	mean (CTRL IN mean)
		standard deviation (CTRL IN std)
		phase drift (CTRL IN phase drift)
	output	mean (CTRL OUT mean)
		standard deviation (CTRL OUT std)
		phase drift (CTRL OUT phase drift)

4.1. Elimination of Long-Term Linear Correlations

Table 2 shows the Pearson correlations (PC) of the components of the optical synchronization system to the influencing components that have a value ≥ 0.8. The standard deviation of the MLO CTRL input has a strong correlation with the air pressure. The SLO also has a strong correlation with the MLO. This makes sense, as the SLO repeats the synchronization signal of the MLO and therefore also all possible fluctuations. These correlations can also be seen in the dependency graph. The other dependencies, which result from the dependency graph, are not used for the elimination of long-term correlations, as the respective Pearson correlations are below the threshold.

Table 2. Correlations before and after trend elimination

Component	Data Channel	Influencing Signal	PC before Isolation	PC after Isolation
MLO	IN stdev	air pressure	0.8521	−0.1005
SLO	IN stdev	MLO IN std	0.8356	0.2874

Main Laser Oscillator

According to the dependency graph, the MLO is directly dependent on the behavior of the MO as well as environmental influences and human-induced disturbances. However, since only data from humidity, temperature and air pressure are available, only these are used to identify possible correlations and subsequently isolate the MLO behavior. Figure 8 shows part of the standard deviation controller input signal compared with the air pressure and the cleaned controller input signal. It is evident to see that the trend of the air pressure has been removed from the controller input. This is confirmed by the fact that the Pearson correlation between the adjusted controller input and the air pressure is -0.1005 and thus much smaller than before the removal.

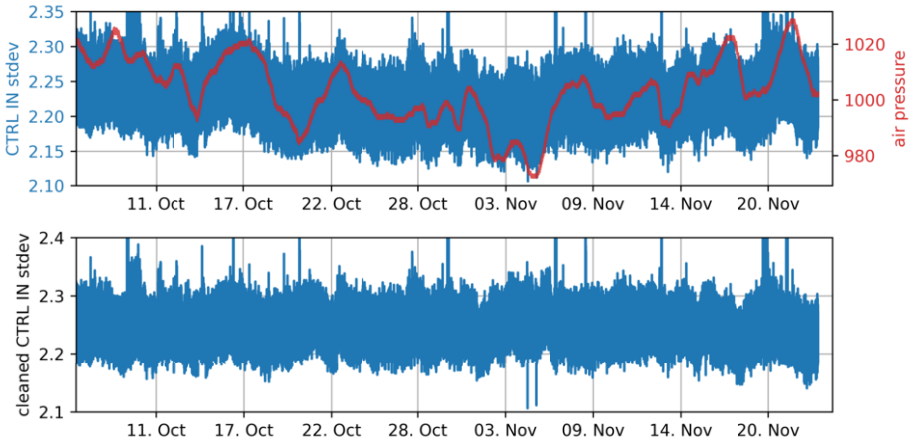


Figure 8. Cleaning the MLO

Link Stabilization Unit

There is no strong correlation between the LSU signals and other influencing components. The greatest correlation of the LSU to other components is the standard deviation of the controller input of the LSU to the mean of the MLO controller output. Because the Pearson correlation of 0.5917 is below the threshold of 0.8, the LSU controller input was not cleaned from the MLO controller output trend.

Sub-Distribution Laser Oscillator

As shown in the dependency graph, the SLO has the environmental influences, the MLO and the LSU as influencing components. In particular, the standard deviation of the SLO controller input shows a strong correlation to the standard deviation of the MLO controller input with a pearson correlation of 0.8356. The trend elimination is shown in Figure 9.

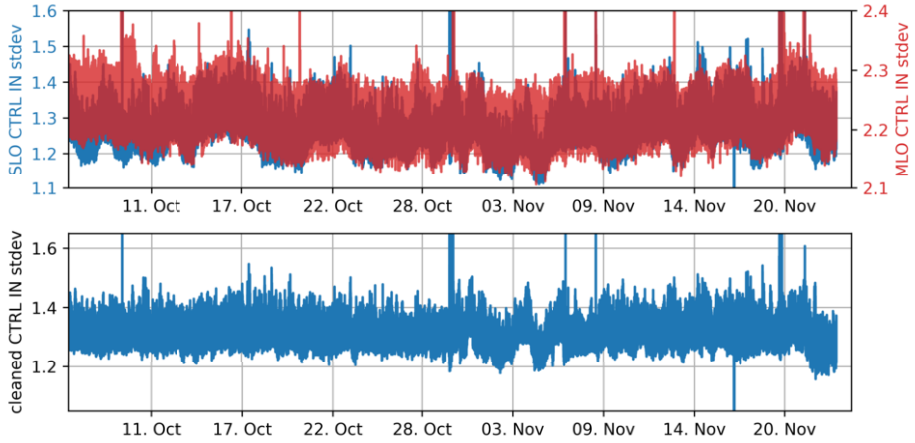


Figure 9. Cleaning the SLO

4.2. Fault Localization and Characterization

In this section the results of the fault localization as described in Section 3.4 are discussed. For each component of the optical synchronization system, it is analyzed how many anomalies were detected by which data channel. Furthermore, for each signal of a system component, it is checked whether a signal of an influencing component also had an anomaly at the same time. This would indicate that anomalous behaviour in the respective analyzed system component did not come from the component itself, but from an influencing component, either another component of the optical synchronization system or environmental influence. The results are summarized in Tables 3, 4, and 5. The anomaly detection described in Section 3.4 results in time ranges of anomalous behavior for all available data channels. Figure 10 shows all environmental data channels and the associated anomaly ranges. Figure 11 shows the MLO data channels and their anomaly ranges, Figure 12 shows the anomaly ranges of the LSU data channels and Figure 13 shows the anomaly ranges of the SLO data channels.

In the following, we have counted how often the time ranges of the analyzed data channel correspond to those in influencing components. As the time ranges of the anomalies sometimes vary greatly between a few seconds and several days, we also made sure that the respective anomalies lasted approximately the same period of time. The Tables 3, 4, and 5 provide overviews of the respective system components, how many anomalies in the data channels of the system components are also anomalies in the influencing data channels and how many anomalies are exclusively in the data channels of the specific component.

Main Laser Oscillator

The results show that relationships to the influencing data channels can be found primarily on the controller input data channels. It can be seen that most of the anomalies of the MLO channels cannot be explained by one of the available influencing channels. In case of the MLO anomalies for which the influencing signals show simultaneous anomalies, it can be seen that these are mainly increases in temperature and humidity. These anoma-

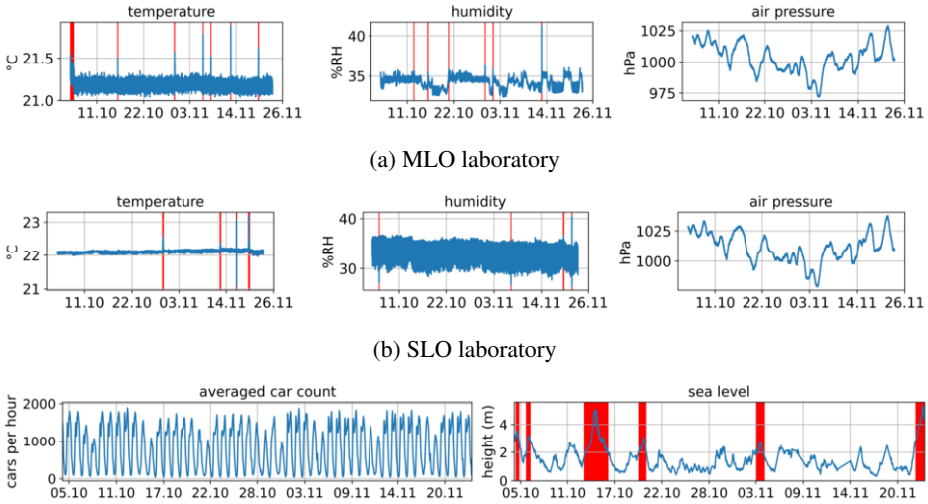


Figure 10. Anomaly ranges of environmental data channels

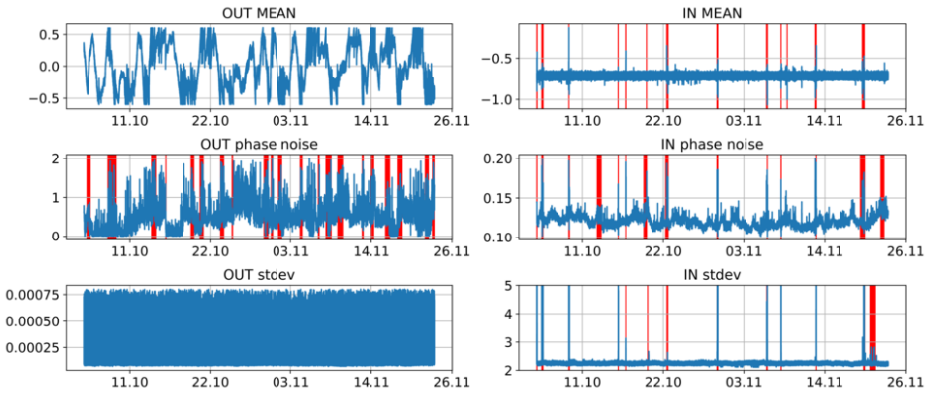


Figure 11. Anomaly ranges of MLO data channels

lies occur particularly on days when the optical synchronization system is undergoing maintenance. The anomalies on the MLO, which occur at the same time as anomalies in the influencing data channels, are therefore more likely to be explained by maintenance work on the system; the changes in temperature and humidity are more likely to be other side effects of these maintenance activities.

Link Stabilizing Unit

The results of the LSU anomalies are shown in Table 4. No anomalous behavior was identified in the data channels of the mean and standard deviation of the controller output. All anomalies found in the controller input standard deviation were also seen in the MLO. This is due to the fact that the MLO phase noise, which is equivalent to the controller

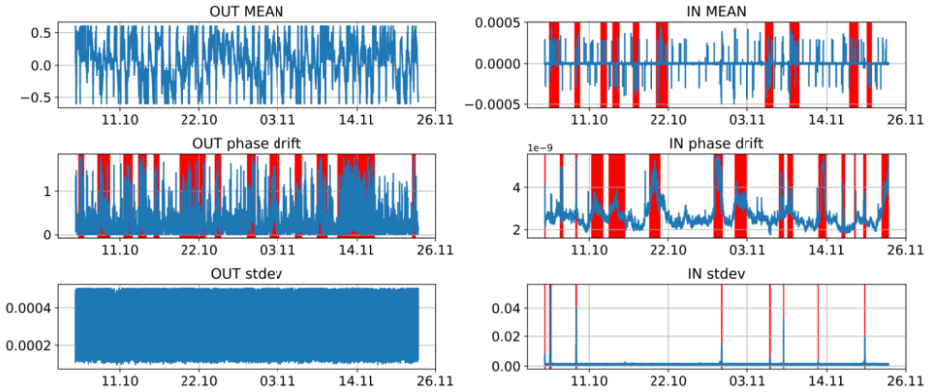


Figure 12. Anomaly ranges of LSU data channels

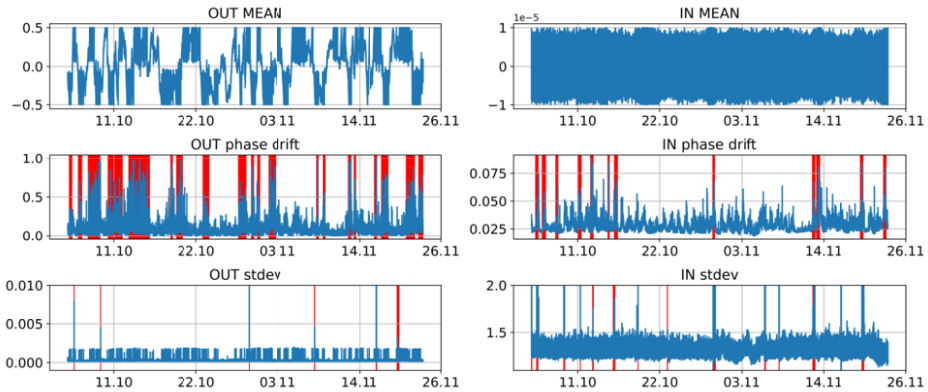


Figure 13. Anomaly ranges (red) of SLO data channels

Table 3. MLO fault localization results

MLO anomalies	CTRL IN mean	CTRL IN stdev	CTRL IN phase drift	CTRL OUT mean	CTRL OUT stdev	CTRL OUT phase drift
total	27	31	13	0	0	20
Exclusive	25	21	10	0	0	17
Air Pressure	0	0	0	0	0	0
Humidity	2	3	1	0	0	3
Temperature	2	10	3	0	0	3

input standard deviation, propagates through the LSU and also serves as input to the LSU controller. The anomalies detected in the LSU controller output can be split into time ranges of few seconds, few hours, and time ranges of several days. The anomalies, which last a few hours, are usually shortly after recorded earthquakes. Most anomalies lasting

several days were also identified in the sea level data. The shorter anomalies were only detected exclusively in the mean of the controller output.

Table 4. LSU fault localization results

LSU anomalies	CTRL IN mean	CTRL IN stdev	CTRL IN phase drift	CTRL OUT mean	CTRL OUT stdev	CTRL OUT phase drift
total	11	35	31	0	0	12
Exclusive	6	0	10	0	0	5
Seismic	1	0	1	0	0	3
Traffic	0	0	0	0	0	0
MLO	4	35	20	0	0	4

Sub-Distribution Laser Oscillator

The results of the fault localization from the SLO are shown in Table 5. For all SLO data channels, the same anomalies are localized in the MLO as in the LSU. This is because phase noise that already occurs in the MLO propagates through the LSU to the SLO. Exclusive anomalies were discovered in particular in the phase noise of the SLO controller output.

Table 5. SLO fault localization results

SLO anomalies	CTRL IN mean	CTRL IN stdev	CTRL IN phase drift	CTRL OUT mean	CTRL OUT stdev	CTRL OUT phase drift
total	0	2	24	16	10	19
Exclusive	0	0	4	5	3	16
Air Pressure	0	0	0	0	0	0
Humidity	0	0	0	0	0	1
Temperature	0	0	0	2	0	2
MLO	0	2	20	9	7	1
LSU	0	2	20	9	7	1

5. Conclusion

The dependency graph is a very simple but powerful method of modeling a complex distributed system and representing dependencies between system components or external influences on the system. Furthermore, the dependency graph offers the possibility to add further components very easily. This is extremely important, as currently only the main components of the optical synchronization system have been considered and even with these additional data channels can be added.

With the help of correlation elimination, slow linear dependencies between the components were eliminated. However, this method does not manage to eliminate spontaneous events that propagate through the distributed system. In addition, there is not necessarily a linear relationship between the system components. It is therefore not sufficient to calculate a linear correlation at this point. However, if there is a linear relationship, for example between the air pressure and the standard deviation channel of the MLO, the trend of the air pressure can be successfully removed.

Component behaviour that disturbs the phase of the synchronization signal is passed on to the subsequent components. This explains why the anomalies detected in the standard deviation of the respective controller inputs also occur in the subsequent system components. Of the external environmental influences, air pressure in particular has an effect on the laser oscillators while seismic activity, which changes the synchronized path length, has an effect on the LSU. The disturbances caused by traffic could not be identified in the analyzed data channels by the presented methods. This is due to the fact that such human disturbances only have a negligible influence on the synchronized distance, especially compared to ocean waves and earthquakes effects.

In conclusion, it can be said that the implemented methods have led to an improved understanding of how certain behaviour is propagated through the optical synchronization system. Due to the ability of the dependency graph to be easily extendable, future work will focus on further populating the dependency graph to better understand the system behavior. Through fault localization, it is now also possible to look at the behavior of individual components in isolation from each other and apply predictive methods to predict future behavior.

References

[1] Sobolev E, Zolotarev S, Giewekemeyer K, Bielecki J, Okamoto K, Reddy HK, et al. Megahertz single-particle imaging at the European XFEL. Communications Physics. 2020;3(1):97.

[2] Heuer M. Identification and control of the laser-based synchronization system for the European X-ray Free Electron Laser [doctoral Thesis]. Technische Universität Hamburg-Harburg; 2018.

[3] Lamb T, Czwalinna M, Felber M, Gerth C, Kozak T, Müller J, et al. Large-scale optical synchronization system of the European XFEL with femtosecond precision. Proc IPAC'19. 2019:3835-8.

[4] Gao Z, Cecati C, Ding SX. A Survey of Fault Diagnosis and Fault-Tolerant Techniques—Part I: Fault Diagnosis With Model-Based and Signal-Based Approaches. IEEE Transactions on Industrial Electronics. 2015;62(6):3757-67.

[5] Gao Z, Cecati C, Ding SX. A Survey of Fault Diagnosis and Fault-Tolerant Techniques—Part II: Fault Diagnosis With Knowledge-Based and Hybrid/Active Approaches. IEEE Transactions on Industrial Electronics. 2015;62(6):3768-74.

[6] He M, Zhang J. A Dependency Graph Approach for Fault Detection and Localization Towards Secure Smart Grid. IEEE Transactions on Smart Grid. 2011;2(2):342-51.

[7] Qi G, Yao L, Uzunov AV. Fault Detection and Localization in Distributed Systems Using Recurrent Convolutional Neural Networks. In: Cong G, Peng WC, Zhang WE, Li C, Sun A, editors. Advanced Data Mining and Applications. Cham: Springer International Publishing; 2017. p. 33-48.

[8] Anceaume E, Le Merrer E, Ludinard R, Sericola B, Straub G. FixMe: A Self-organizing Isolated Anomaly Detection Architecture for Large Scale Distributed Systems. In: Baldoni R, Flocchini P, Binoy R, editors. Principles of Distributed Systems. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. p. 1-15.

[9] Nawaz A, né Hoffmann CH, Graßhoff J, Pfeiffer S, Lichtenberg G, Rostalski P. Probabilistic model-based fault diagnosis for the cavities of the European XFEL. at - Automatisierungstechnik. 2021;69(6):538-49. Available from: https://doi.org/10.1515/auto-2020-0064.

[10] Schulz S, Grguraš I, Behrens C, Bromberger H, Costello J, Czwalinna MK, et al. Femtosecond all-optical synchronization of an X-ray free-electron laser. Nature communications. 2015;6(1):5938.
[11] Grünhagen A, Eichler A, Tropmann-Frick M, Fey G. Data-Based Condition Monitoring and Disturbance Classification in Actively Controlled Laser Oscillators. In: Information Modelling and Knowledge Bases XXXV. IOS Press; 2024. p. 94-114.
[12] Murray JR, Svarc J. Global Positioning System Data Collection, Processing, and Analysis Conducted by the U.S. Geological Survey Earthquake Hazards Program. Seismological Research Letters. 2017 03;88(3):916-25. Available from: https://doi.org/10.1785/0220160204.
[13] Verkehrsstärken Hamburg. Landesbetrieb Geoinformation und Vermessung; 2023. Available from: https://www.hamburg.de/bsw/landesbetrieb-geoinformation-und-vermessung/.
[14] Seegangsportal. Bundesamt für Seeschifffahrt und Hydrographie; 2023. Available from: seastate.bsh.de/.
[15] Ramm F, Topf J, Chilton S. OpenStreetMap. Die freie Weltkarte nutzen und mitgestalten. 2010;3:978-3865413758.
[16] Pearson K. VII. Note on regression and inheritance in the case of two parents. proceedings of the royal society of London. 1895;58(347-352):240-2.
[17] Jeong YS, Jeong MK, Omitaomu OA. Weighted dynamic time warping for time series classification. Pattern Recognition. 2011;44(9):2231-40. Computer Analysis of Images and Patterns. Available from: https://www.sciencedirect.com/science/article/pii/S003132031000484X.
[18] Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition. IEEE transactions on acoustics, speech, and signal processing. 1978;26(1):43-9.
[19] Breunig MM, Kriegel HP, Ng RT, Sander J. LOF: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD international conference on Management of data; 2000. p. 93-104.