Performance Variability and Causality in Complex Systems

Anwesha Das[1] Daniel Ratner[2] Alex Aiken[1]
[1]*Stanford University* [2]*SLAC National Accelerator Laboratory*

***Abstract*—Anomalous behaviour in subsystems of complex machines often affect overall performance even without failures. We devise unsupervised methods to detect times with degraded performance, and localize correlated signals, evaluated on a system with over 4000 monitored signals. From incidents comprising both downtimes and degraded performance, our approach localizes relevant signals within 1.2% of the parameter space.**

***Index Terms*—Time-series, Causal Analysis, Degradation**

I. Introduction

Despite previous efforts examining performance degradation it is still challenging to diagnose problematic subsystems resulting in unsteady machine performance [29]. We consider cyberphysical systems (CPSs) such as *light sources*. X-ray lasers are a kind of light source generating high energy beams for scientific experiments. Light sources are complex with multiple heterogeneous components (e.g., Radio-Frequency (RF) cavities [24], magnets, water systems), and even during normal operation, subsystem behaviour varies with time and physical location making aberrations hard to capture. Identifying malfunctions from among the considered signals of the system (which we call the *search space*) while the machine is operational is non-trivial. Existing threshold-based alarms and monitoring tools are inadequate to quantify performance degradation in its early-stages, leaving less time for interventions before complete failure occurs.

We propose automated methods to localize a subset of signals related to *causal subsystems*; i.e., subsystems related to a performance anomaly. Our goal is not to pinpoint the root cause, but to find unusual signal trends that can be a side-effect or cause of anomalous behaviour, guiding operators in diagnosing the problem. Quantifying degradation and causality has two benefits. First, the subsystem specialist can potentially fix problems before a failure happens. Second, even in the absence of faults, unexpected signal correlations can bring to light abnormalities in their infancy, aiding preventive maintenance.

Existing studies on performance diagnosis [17, 28] have investigated approaches to enhance system reliability. As subsystem functioning can be complicated enough to require dedicated investigation [10], most studies on physical systems analyze subsystems separately (e.g., magnet power supplies [15], heat pump [16]) without considering overall machine performance. Some studies rely on injected faults [9], or assume labeled data [24]. We develop an approach for causal analysis of performance degradation on real production data from entire systems. We exploit the following observations:

a) Using multiple performance-indicative signals helps capture important correlations. (§ II)

b) All signals of a subsystem may not have similar trends, or correlate with machine performance to the same extent. Conversely, signals from multiple subsystems can collectively hint at anomalous conditions. (§ II)

c) Signals can be inherently noisy (i.e., possess high variability) whether or not they are anomalous. Examining a signal's general behaviour (compared to several time-windows) can help identify non-causal signals. (§ IV)

d) Knowledge of a CPS's physical architecture helps identify signals that are causes rather than effects of problems. (§ II)

Based on these insights we make the following contributions:

1. We propose a method to classify machine degradation.
2. We perform unsupervised causal analysis to shortlist a subset of signals pertaining to certain subsystems related to anomalies.
3. We evaluate our approach using cases from a production CPS. Our approach achieves below 15% errors in detecting degraded performance, and ranks relevant signals within the top 1.2% ($\leq$50) of the considered set of signals.

II. Background and Motivation

Particle accelerators such as the X-ray free electron lasers and storage rings generate high-energy beams [1]–[3]. Beam experiments are used for a variety of activities ranging from soft error assessment in memory chips to probing atomic properties of materials [23]. These are large systems, often kilometers in length, with many subsystems and thousands of sensors that generate multivariate time-series data used for real-time monitoring. Subsystem problems result in beam downtime (i.e., system failure) or intermittent beam performance degradation, preventing effective utilization of an expensive machine. Currently, fault repair happens after a subsystem fault manifests. For transient beam instabilities, engineers hand pick a few signals based on experience to assess what may have caused unexpected jitter at a specific location. Such manual efforts are prone to missing subtle correlations and may come too late to catch problems before a complete failure.

Most systems have a few metrics (e.g., so-called *golden signals* [4]) that are used for analyzing overall performance. Figures 1 and 2 show a degraded beam using two such golden signals, namely, *beam intensity* and *pulse energy*. The *x*-axis shows timestamps and the *y*-axis shows the normalized values from the chosen signals. While *intensity* improves around hour 6, *pulse energy* remains low during the same time (Fig. 1). In Fig. 2, both the signals drop around hour 13:30 indicating an

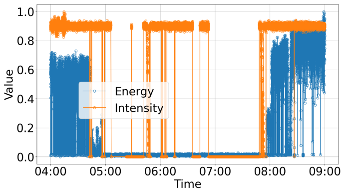

Fig. 1: System Performance

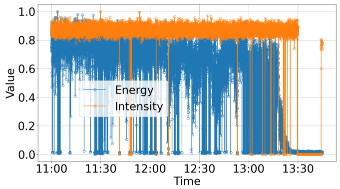

Fig. 2: Degradation

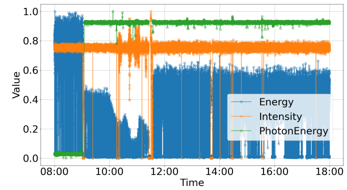

Fig. 3: Parameter Tuning

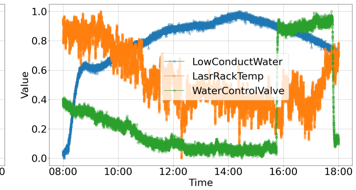

Fig. 4: Subsystem Aberrations

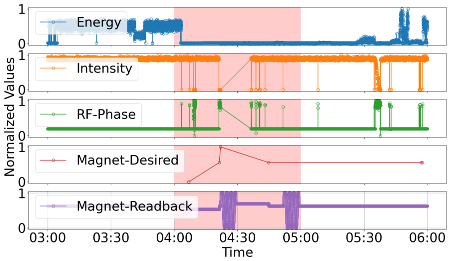

Fig. 5: Correlation vs. Causation

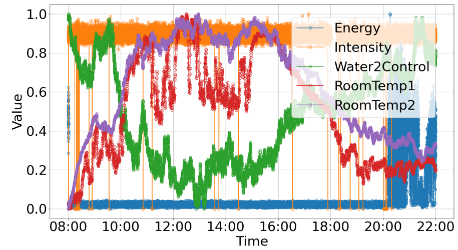

Fig. 6: Subsystem Correlations

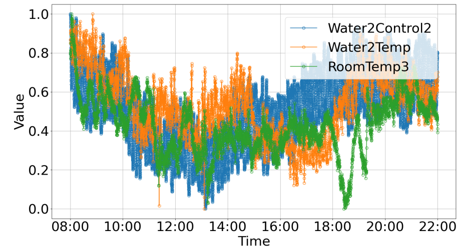

Fig. 7: Subsystem Signals

anomalous condition, while lack of major variations from the 11^{th} to 13^{th} hours imply normalcy.

Challenges: The major hurdles for anomaly diagnosis are: **Tuning side-effects**: User experiments require specific machine configurations achieved via manual parameter tuning. Such parameters, called *control variables*, can trigger changes in other signals, including performance-indicative ones. Intentional signal adjustments produce time-series deviations similar to anomalous fluctuations. Reconfigurations in the face of unforeseen subsystem glitches, or environmental variations beyond our control, add to the machine's dynamic nature. Assessing normal patterns and anomalies amidst time-varying machine states is non-trivial. We track changes in the control variables and analyze signal correlations when the machine should be stable in the absence of major adjustments. (§ IV)

Figures 3 and 4 give an example of degradation that involves tuning with 6 signals: 2 related to performance, 1 a tuned parameter (photon energy), and the remaining 3 related to a laser and water system. A performance drop is discernible from hour 9 to 11:35 through the pulse energy signal, which coincides with a manual increase of photon energy at hour 9 (Fig. 3). Fig. 4 shows that the laser room temperature drops with this performance drop, indicating an anomalous condition. The water signals close to the laser system show variations for the next few hours. It is not clear if the water system contributes to the degradation. In this case, the intentional photon energy change caused the performance drop, and thus is not anomalous. In the absence of subsystem correlations, a performance drop can be a temporary side-effect of a manual adjustment. Such simultaneous events make causal analysis difficult. Knowing such subtle yet important signal correlations, if feasible, can be helpful for diagnosing faulty subsystem behaviour.
Localization accuracy: Our goal is to produce a short list of signals that are related to the cause of machine degradation without discarding highly relevant signals. Many false positive signals (i.e., unrelated signals that appear causal) in a short list (e.g., 25-30), or causal signals buried in a large list (e.g., 200-300) can hinder timely diagnosis. Signals from fewer areas, over subsystems from a set of geographically dispersed signals can be more useful in inferring the main anomalous subsystem. Univariate signal trends are further analyzed to obtain a signal list that is short yet precise. (§ IV)

Signals exhibiting variation during an episode of performance degradation are not necessarily causal; what we are looking for is unusual variability. Certain signals fluctuate during performance variations no matter what the anomalous subsystem is. Figure 5 shows an example to highlight that a highly correlated signal need not be causal, with signal variations between 3:00 and 6:00 hours. The performance drops at ≈4:00 hours, discernible through the energy and intensity signals. A specific RF-phase signal shows strong correlation to performance, however, the RF-system is not the cause, but a side-effect. The magnet-related signals are weakly correlated, yet are closer to the cause: an unexpected jump in the magnetic field induces an anomalous drop in performance.
Lack of time-delay: Signals vary across distant locations with insignificant propagation delays. Establishing a *happens-before* relation [6] is inappropriate in a rapidly changing system where concurrent events are common and subsystem interactions are not fully known. We try to eliminate non-causal signals without forming any inter-signal relationships. (§ IV)

Systemic Guidance: The machine mostly has a linear structure with major subsystems arranged roughly in a pipeline. The beam is delivered downstream, at the experimental halls. When multiple signals are correlated with an anomaly, signals that are more upstream are more likely to be causes than effects. Apart from control variables, subsystems with feedback control can have non-causal signals that may not hint at anomalies (e.g., supply temp.). Other CPSs with a non-linear structure can also provide architectural guidance that can help eliminate non-causal signals (e.g., regions closer to beam delivery, the control vs. physical layer in industrial control systems [8]).

Figure 6 shows an example of degraded performance that correlates with a water system's control valve variations. Co-located with this water system, air temperature signals for the room (which houses electrical equipment) exhibit anomalous behaviour, suggesting a good starting point for manual inspection. Some other signals of the same subsystems do not show strong correlations, and appear unrelated, as seen in Figure 7. Such relatively stable signals of problematic subsystems do not help with causal evidence.

Performance degradation is defined as increased variability of golden signals relative to a reference time of normal behaviour. Not all variations can be explained by the available signals. Certain control software problems, for example, may not have any underlying anomalous symptoms. Nonetheless, capturing those cases where knowing the defective subsystem can reduce recovery time can be beneficial.

III. Related Work

Performance variability has been researched for supercomputers, clouds, the Internet, and other software service systems using workload characteristics and resource usage metrics [7, 12, 14, 25, 29] through information theoretic- [11, 14], Machine Learning-based [25, 28], and statistical methods [7, 20, 25]. Some predict [29] performance, while others analyze variability [12]. Workload-aware studies [25] are not relevant in our context, as we do not have access to experiment meta-data. The challenge of dealing with manual parameter setting is not explored in some prior studies [7, 12]. While the nature of anomalies (e.g., memory leaks, network jitter) studied in the past [17, 18] is quite different from ours (e.g., abrupt RF-phase jumps), they affirm the importance of integrating domain insights (e.g., operator feedback) into the solution for improved accuracy.

Many studies [6, 14, 17, 18, 22, 25, 28] conduct fault localization, and anomaly diagnosis using hypothesis testing, deep learning, feature aggregation, decision trees, and fault injection. In contrast to these, we use fine-grained filtering by examining signal trends to localize the search space, in the absence of any synthetic data. CPSs often experience multiple spatially distant anomalous events with negligible time delays for which, graph or relation-based methods [21, 22, 26] may be less viable for timely problem diagnosis.

Some related studies [8, 11, 19] on CPSs (e.g., X-ray lasers, microgrids, electrical plants) perform subsystem-oriented fault analysis that can complement our work. Past approaches include isolation forests, graphs, clustering, and Bayesian models [5]. Regression models that work for IT service systems may not be effective for physical systems [22] based on the quality of time-series data. Additional methods analyzing signal correlations, such as ours, can potentially further benefit the community.

IV. Approach Overview

Complex systems often have no automated way to determine degradation [10]. A few signal spikes and dips may be caught visually, however, a systematic approach to identify degradation with potential anomalies can help human-in-the-loop CPSs.

Degradation Detection: Performance is classified into three categories, a) `OK`, b) `Unsteady`, and c) `Degraded`. `OK` implies acceptable performance, `Unsteady` indicates moderate variability that can be due to anomalous conditions, and `Degraded` implies deteriorating performance. `Degraded` suggests a relatively higher degree of unstable performance over `Unsteady`. Downtime or failure is a special case of degradation when the system has halted. Algorithm 1 shows the major steps of our method. For a given system, the performance-indicative signals, (i.e., sig_{perf}), and the commonly used control variables, (i.e., sig_{con}) are first identified. The latter are often used to configure the system. As low signal values are often anomalous in our context, minimum values seen in specific time-windows (T_W) are used to replace any missing values to retain temporal relevance. Irregular signals (unevenly spaced) are normalized and regularized.

Algorithm 1 Detection of Performance Degradation

Require: T_W, sig_{perf}=[P_1, P_2,... P_k], sig_{con}=[C_1,... C_m], α, β
Ensure: P_{class} ▷ Label Performance Quality
procedure Detect Variability(T_W, sig_{perf}, sig_{con}, α, β)
  [$\text{sig}_{perf'}$, $\text{sig}_{con'}$] ← Normalize (sig_{perf}, sig_{con})
  $\text{sig}_P \leftarrow \sum_{P_1}^{P_k} \text{sig}_{perf'}$, Representative signal for performance
  [CP_{perf}] ← detect ChangePoint ($\text{sig}_P{}^{T_W}$, size)
  [$CP_{control}$] ← detect ChangePoint ($\text{sig}_{con'}{}^{T_W}$, size)
  if ($CP_{control} > \alpha$), $C_i \in \text{sig}_{con}$ **then**
    Ignore Variability ▷ Manual adjustments
  else ▷ Control variables do not change
    if (CP_{perf} = 0) **then** P_{class} ← OK **else**
    if ($CP_{perf} \leq \beta$) **then** P_{class} ← Unsteady **else** P_{class} ← Degraded
return P_{class} ▷ Category

As all performance-indicative signals may not fluctuate to the same extent but any can be indicative of degradation (e.g., Fig. 1), we combine the performance-indicative signals into a single derived signal. We apply prior work [13] on breakout detection to identify *change points* on the derived signal (i.e, sig_P). Timestamps that maximize the difference between the statistical characteristic of adjacent sub-patterns of a time-series are estimated as *breakouts* or *change points*. The idea is to track changes in both control variables and the performance signal. In case of a major change in one or more control variables, any noticeable performance degradation is less likely to be anomalous. We measure change points in the control variables during normal times devoid of performance degradation or manual adjustments, to identify a threshold characteristic of normal behaviour (i.e., α). This threshold is compared with the number of change points observed while examining degraded performance. When control variables are steady, based on the distribution of the number of change points, we classify the level of degradation. For minimal variability indicated by 0 change points, performance is labeled as `OK`. If the number of change points is below a threshold (e.g., β =4) the time-window is tagged as `Unsteady`, otherwise `Degraded`.

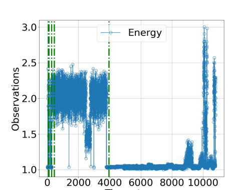

Fig. 8: Energy

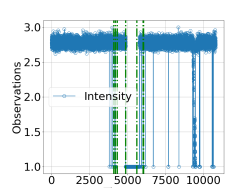

Fig. 9: Intensity

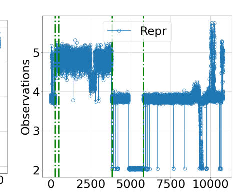

Fig. 10: Combined

Figures 8 and 9 show energy and intensity signals, with 6 and 8 change points (i.e., CPs), respectively. Figure 10 shows that

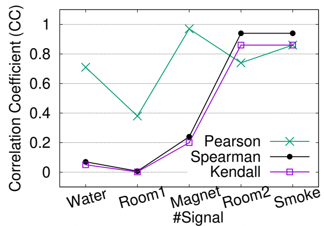

Fig. 11: Correlation

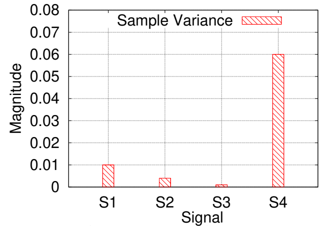

Fig. 12: Variance

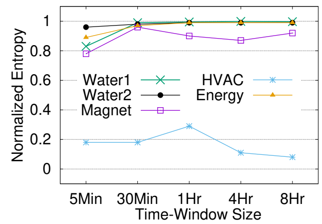

Fig. 13: Diverse Durations

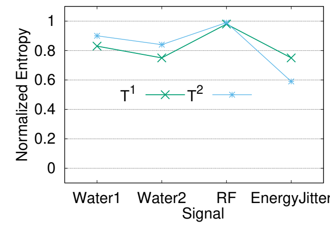

Fig. 14: Signal Entropy

the combined signal has 4 change points, classifying this case as `Unsteady`. The lower energy values, besides intensity drops at ≈7000, and ≈9500 timesteps respectively, are inadequate to count as change points in the combined signal. The number of change points in the resulting signal need not coincide with the min, max or average of the individual signals' change points. Sustained poor performance in the absence of variability (i.e., few or no CPs) due to subsystem or system failures may not be identified as `Degraded` by Algo. 1, as our method is designed towards capturing short-term variability.

The `size` parameter indicates the number of observations used for detecting change points. We find that 20 is a suitable `size`, such that minor signal deviations are captured while redundant change points during normal signal drops are reduced. Note that for more stable systems *size* would be selected to be longer, for more dynamic systems it would be shorter.

Algorithm 2 Causal Analysis

Require: T_W, sig$_{perf}$=[P$_1$,..P$_k$], sig$_{con}$=[C$_1$,..C$_m$], sig$_L$=[S$_1$,..S$_N$]
Ensure: sig$_{reduced}$=[S$_{11}$, S$_{26}$, S$_{37}$...] ▷ Shortlisted Signals
procedure SEARCH SPACE REDUCTION(T_W, sig$_{perf}$, sig$_{con}$, sig$_L$)
  [sig$_{L'}$] ← sig$_L$ \ $\forall S_i \in$ sig$_{con}$, T_W ▷ Adjusted signals
  [sig$_L^{corr}$] ← Bivariate Correlation (sig$_{L'}$, sig$_{perf}$)
  [sig$_{denoise}$] ← Noise Filter (sig$_L^{corr}$)
  [sig$_{En}$] ← Signal Entropy (sig$_{denoise}$)
  [sig$_{Short}$] ← Dynamic Modeling (sig$_{En}$)
  Score^{S_i} ← $\frac{(W_1 * CC^{S_i}) + (W_2 * Var^{S_i}) + (W_3 * E_P^{S_i}) + (W_4 * Comp^{S_i})}{\sum W_j, 1 \le j \le 4}$
  [sig$_{Short}$] ← Rank S$_i$ based on Score
  [sig$_{reduced}$] ← sig$_{Short}$ ≅ System-Guided Ordering
return sig$_{reduced}$ ▷ Localized Search Space
procedure BIVARIATE CORRELATION(sig$_{L'}$, sig$_{perf}$)
  [CC$_P^{S_i}$, CC$_S^{S_i}$, CC$_K^{S_i}$] ← Correlate(S$_i$, sig$_{perf}$), $\forall S_i \in$ sig$_{L'}$
  CC^{S_i} ← max [$|CC_P^{S_i}|$, $|CC_S^{S_i}|$, $|CC_K^{S_i}|$], $\forall S_i \in$ sig$_{L'}$
  sig$_L^{corr}$ ← sig$_{L'}$ \ ($CC^{S_i} \le \gamma$), $\forall S_i \in$ sig$_{L'}$ ▷ Threshold γ
return sig$_L^{corr}$
procedure NOISE FILTER(sig$_L^{corr}$)
  [Var^{S_i}...] ← [T_W^j], $\forall S_i \in$ sig$_L^{corr}$, j $\le M$ ▷ M Times
  [sig$_{denoise}$] ← sig$_L^{corr}$ \ [Var$^{S_i} \ge \theta$], ∀ S$_i$ ▷ Threshold θ
return sig$_{denoise}$ ▷ Noise Filter
procedure SIGNAL ENTROPY(sig$_{denoise}$)
  [E$_P^{S_i}$...] ← Compute E$_P$ ∀ S$_i$ ∈ sig$_{denoise}$
  [sig$_{En}$] ← sig$_{denoise}$ \ [E$_P^{S_i} \le \delta$], ∀ S$_i$ ▷ Threshold δ
return sig$_{En}$ ▷ Low pattern evolution
procedure DYNAMIC MODELING(sig$_{En}$)
  [Comp^{S_i}...] ← Univariate Trends ∀ S$_i$ ∈ sig$_{En}$
  [sig$_{Short}$] ← sig$_{En}$ \ [Comp$^{S_i} \le \beta$], ∀ S$_i$ ▷ Threshold β
return sig$_{Short}$ ▷ Lower Trends

Causal Analysis: Algorithm 2 summarizes our approach to filtering signals. The relationship between each signal and system performance is examined using Spearman's, Pearson's, and Kendall's correlation coefficient (CC ∈ ±1) for a specific time-window (T_W). For a given system, a single performance metric or a combined representative signal can be used for computing correlation coefficients. Spearman's correlation captures monotonic relationships better while Pearson's is suitable for linear associations. Kendall's and Spearman's correlation strengths are often similar. The intuition is to first develop a coarse-grained filter to exclude signals that do not deviate in close temporal proximity to machine performance. The maximum absolute value of CC among the three correlations is used for each signal, to form a ranked list with decreasing CC. Signals with weak correlations (i.e., CC $\simeq$ 0, based on a threshold, e.g., $\gamma = 0.09$) are removed. Figure 11 shows the CCs of 5 signals, 2 related to rooms housing electrical devices, and 1 each from a smoke detector, water-system, and magnet, respectively, over a 4-hour time-window. Though Pearson's correlation is higher for the magnet signal (0.97), it may not be causal. Water and Room signals have lower correlations (≤0.86), yet, are helpful in detecting anomalies.

To identify signals with inherent high variability, their mean and variance are compared across multiple time-frames. Signals that frequently vary during operational times are either noisy or relate to commonly affected non-causal subsystems. Figure 12 shows 4 signals, S1 to S4, where S4 has higher variance (0.06). Signals with variance above a threshold (e.g., $\theta = 0.02$) are removed, as it is difficult to determine whether they are causal or not in diagnosing subsystem defects.

We further assess signal behaviour using permutation entropy (E_P) [11], which is more noise resistant in chaotic time-series.

$$E_P = -\sum p(\pi_j) log_2 p(\pi_j) \tag{1}$$

A time-series is partitioned into multiple sub-vectors using the embedding dimension and time-delay hyperparameters. An ordinal pattern (π_j) is one of the permutation patterns of these sub-vectors. E_P is Shannon's entropy [11] over the probability distribution of the ordinal patterns ($p(\pi_j)$), for any time-window. Signals with large entropy indicating higher complexity are retained, and those below a threshold (i.e. δ) are dropped. The aim is to measure the relative dynamical change of patterns emerging in signals. Figure 13 shows the entropy values for 5 signals at 5 different time-spans. As seen, E_P need not be proportional to the window size. While causal signals (e.g., water) can score high, helpful signals (e.g., HVAC, a heating unit) can score low if the sampling rate is low. Figure 14 shows that for time-frames T^1 and T^2, E_P correctly scored the non-causal signal (i.e., EnergyJitter) lower than the others. Signals that may have been averaged with loss of resolution often get filtered in this step.

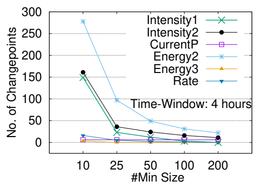

Fig. 15: Change Points

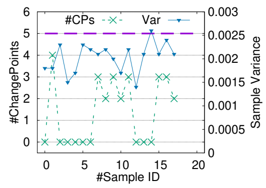

Fig. 16: Performance

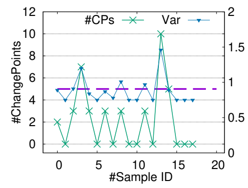

Fig. 17: Categories

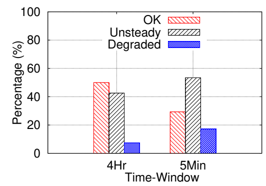

Fig. 18: Classification

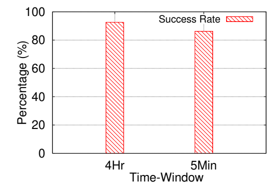

Fig. 19: Success Rate

To detect finer time-varying changes in the signals, a dynamic linear model is used. Dynamic models [27] are used for statistical analysis of time-series by modeling the underlying state space. Fig. 20 shows a signal with lower trend that may be less causal compared to a signal exhibiting an interesting pattern around the time of degraded performance, as seen in Fig. 21. In Fig. 20, the filtered, predicted, and smoothed components of a signal, respectively, have low variability. In Fig. 21, a spike is evident around the 2200^{th} time-step in another signal. The trends over short time-bins are measured, and signals with no or negligible trends are filtered. Signals with better trends rank higher after this step.

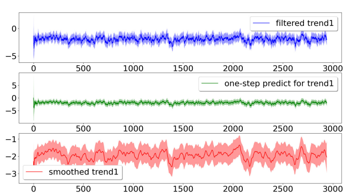

Fig. 20: Linear Model

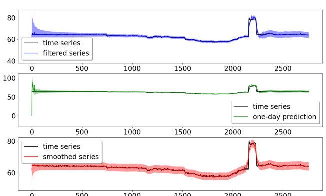

Fig. 21: Univariate Trend

The remaining signals are ranked after assigning a score, based on the weighted average of all the steps. Finally, the ranked signals are ordered based on the machine layout, if feasible. This step can rank a causal signal higher (e.g., based on location) in certain cases. We give more weight to bivariate correlation and entropy to filter non-causal signals. Our weights are experimentally derived, and thresholds are estimated based on the distribution of metrics. Suitable weights can be assigned based on domain insights, or to reflect the causality of signals.

V. Evaluation

Datasets: We evaluate 29 cases for a light source, logging over 4000 to 8000 signals per year, as shown in Table I, related to degraded performance and downtimes. The time-duration of these cases range between 1 and 4 hours. We validate our analysis with engineer logged reports (i.e., ground truth). We

TABLE I: System Details

System	# Signals	Cases	Subsystem
LCLS [2] (LS1)	4184-8934	29	Water, Magnet, RF, Electrical
NSLSII [3] (LS2)	70	8	Water
APS [1] (LS3)	1331	3	Magnet Power Supply

also study 8 and 3 cases from two other machines with 70 and 1331 signals, respectively, to assess the generality of the proposed methods. The anomalous events relate to problems in water systems, magnets, RF cavities, and electrical devices. The three systems are similar, but not identical in terms of the overall architecture and parameter space.

Degradation Detection: System performance is studied with the performance-indicative signals (i.e., intensity and energy) using Algo. 1. Figure 15 shows that for a specific 4-hour time-frame, certain signals with enhanced data resolution reflect variations better (e.g., Intensity2, Energy2) with more change points (CP). As the size parameter of change point detection method is increased from 10 to 200, the number of CPs do not change significantly for certain signals (e.g., rate) indicating lack of major variability. Signals not showing finer variations may be deficient of the necessary frequency and regularity of data needed for analysis. Figures 16 and 17 show the variance and CPs for 18 samples from different years, with the same threshold (i.e., β=5). Fig. 17 shows that the CPs of two samples exceeding the threshold are considered degraded, while the CPs in Fig. 15 remain below the threshold. The magnitude of variance may not be a reliable indicator of degradation, since for the same number of CPs (e.g., 2), the variance can be different (e.g., 0.002 vs. 0.89). Though variance can be slightly higher during degraded performance relative to normal times, it does not adequately estimate the level of degradation.

For LS1, time-windows of 4-hour, and 5-min are studied (we had insufficient data to conduct the same studies for LS2 and LS3). A few hours help capture degradation, while short time-scales help flag fine-grained changes and estimate detection errors. Figure 18 shows that 29.3% to 50% of times are `OK`, 42.6% to 53.4% are `Unsteady`, while below 17.2% are classified as `Degraded` performance. ≈86.2% and 92.6% of the times are found to be correctly classified by our detector (i.e., <15% errors), as seen in Fig. 19. Cases where the statistical nature of normal and anomalous times are similar (leading to errors), are often observed in CPSs [10].

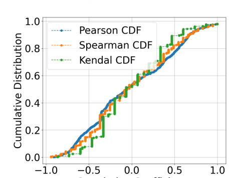

Fig. 22: CDF

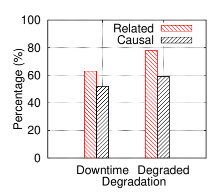

Fig. 23: Related

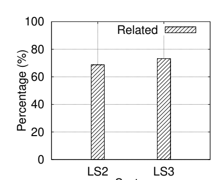

Fig. 24: Accuracy

Causal Analysis: Figure 22 shows the cumulative distribution function (CDF) of the three correlation coefficients calculated to check each signal's association with LS1 performance. The fraction of signals with strong (positive or negative) correlation (i.e., close to ±1) is not high (≤15%) (Fig. 22). However, several signals have moderate to weak correlations (≈ 60% → ±0.5) that need to be further examined for causal analysis. For LS1, cases of downtime (19) and degraded performance (10) are identified. Figure 23 shows that 52% to 59% of the shortlisted signals are *causal*, while 63% to 78% are *related*. Related signals include causal signals as well as those signals affected by the anomalous condition. While it is practically difficult to validate every shortlisted signal, based on the identified primary subsystem problem, the non-causal signals are excluded. As an example, for a water system (W_1) problem, signals related to W_1 are *causal*, and affected signals from the same or near by area as W_1 are *related*. For LS1, the final list of signals comprising correlated signals using Algo. 2 is below 50, which is $\leq$1.2% of the considered signals.

Figure 24 shows that 68.7% to 73.2% of the signals in LS2 and LS3 are related to faulty conditions. As only signals pertaining to a few subsystems are accessible, not all steps of Algo. 2 are applicable. The performance-indicative signals for LS2 were unavailable. For LS3, we had inadequate data pertaining to normal times, to derive the noise filtering threshold. Despite these limitations, some of the related causal signals are correctly identified using our approach.

TABLE II: Related Signals

Water^{Degraded}			**RF**^{Downtime}		
R	**Signal**	**S**	**R**	**Signal**	**S**
2	Wat1-Loc1	68	1	RF-Loc3	64
4	Wat2-Loc1	62	2	RF-Loc2	59
9	Room-Loc2	37	3	Wat1-Loc4	48
11	RF-Loc1	35	5	Wat2-Loc4	43

TABLE III: Rank Improvement

Signal	**Rank**			
	CC	**+N**$_F$	**+E**$_P$	**All**
Wat1-Loc1	50	27	18	2
Room-Loc2	48	39	21	9
RF-Loc2	38	24	16	2
Wat1-Loc4	149	71	26	3

Table. II shows examples of degraded performance due to a faulty water system, and downtime caused by an RF trip, through a few selected signals with their ranks (R) and correlation scores (S). Based on the fraction of subsystem signals evident for a specific location (e.g., 2 water signals from location 4 for RF-caused downtime), engineers can hypothesize potential defects. It helps to limit the physical areas for problem diagnosis, so that even in the presence of non-causal signals, not many areas (usually distant from one another) need inspection. The divergence in correlation scores of signals, if any, can also help prioritize certain signals over others for diagnosis. Table. III shows how the ranks improve after subsequent steps (CC: Correlation Coefficient, N_F: Noise Filter, E_P: Entropy) of causal analysis for a few signals. For certain cases, the related signals are well correlated to performance leading to higher ranks at the outset (e.g., RF-Loc2 ranks 38 for CC), while for those that are weakly-correlated, the ranks improve with successive filtering (e.g., rank 149 to 3 of Wat1-Loc4). Further enhancement using domain insights is subject to future work.

Our design strives to achieve a balance between coarse correlation and finer temporal changes. Our modular approach of filtering can be suitably re-ordered to eliminate non-causal signals and is generally applicable for systems with time-series data spanning both anomalous and normal times. The localization efficacy depends on a system's dynamic nature, besides the sampling rate, regularity, and noise-level of signals. Higher temporal density in time-series generally helps. For cases with bi- or unidirectional cause and effect relationships, small-scale changes (e.g., O(mins)), and discrete-valued signals, pinpointing problem sources with confidence needs more study.

VI. Conclusion

We describe methods to detect performance degradation with search space localization for causal analysis. Our approach achieves over 90% success in detecting degradation and ranks relevant signals among the top 1.2% of the considered signals. Our experimental insights about signal variations has the potential to improve the reliability of diverse complex systems.

Acknowledgment

We are grateful to Sean Kalsi, and other staff engineers for sharing their domain insights, and acknowledge the support received from the Scientific Data Facility for evaluation. This work is supported by LCLS development program, and the U.S. DOE, under contract DE-AC02-76SF00515 at SLAC NAL.

References

[1] "Advanced Photon Source," https://www.aps.anl.gov/About/Overview.
[2] "Linac Coherent Light Source," https://lcls.slac.stanford.edu/.
[3] "National Synchrotron Light Source," https://www.bnl.gov/nsls2/.
[4] P. Aggarwal, S. Nagar *et al.*, "Causal modeling based fault localization in cloud systems using golden signals," in *IEEE CLOUD*, 2021.
[5] V. Aglietti, N. Dhir, J. González, and T. Damoulas, "Dynamic causal Bayesian optimization," in *NeurIPS*, 2021, pp. 10 549–10 560.
[6] P. Chen *et al.*, "CauseInfer: Automated end-to-end performance diagnosis with hierarchical causality graph in cloud environment," *IEEE TSC*, 2019.
[7] D. Duplyakin, A. Uta, A. Maricq, and R. Ricci, "In datacenter performance, the only constant is change," in *IEEE/ACM CCGRID*, 2020.
[8] C. Feng *et al.*, "A systematic framework to generate invariants for anomaly detection in Industrial Control Systems," in *NDSS*, 2019.
[9] C. Feng and P. Tian, "Time series anomaly detection for Cyber-physical systems via neural system identification and Bayesian filtering," in *ACM SIGKDD*, 2021, pp. 2858–2867.
[10] Q. Han, P. Nguyen *et al.*, "Toward an integrated approach to localizing failures in community water networks," in *IEEE ICDCS*, 2017.
[11] Z. Huo, M. Martínez-García *et al.*, "Entropy measures in machine fault diagnosis: Insights and applications," *IEEE Trans. Instrum. Meas.*, 2020.
[12] A. Iosup, N. Yigitbasi, and D. H. J. Epema, "On the performance variability of production cloud services," in *IEEE/ACM CCGrid*, 2011.
[13] N. A. James, A. Kejariwal, and D. S. Matteson, "Leveraging cloud data to mitigate user experience from 'breaking bad'," in *IEEE BigData*, 2016.
[14] M. Jiang *et al.*, "Efficient fault detection and diagnosis in complex software systems with information-theoretic monitoring," *TDSC*, 2011.
[15] J. S. John *et al.*, "Real-time artificial intelligence for accelerator control: A study at the Fermilab booster," *Phys. Rev. Accel. Beams*, 2021.
[16] M. Khosravi *et al.*, "Machine learning-based modeling and controller tuning of a heat pump," in *Journal of Physics: Conference Series*, 2019.
[17] L. Li, X. Zhang, X. Zhao, H. Zhang *et al.*, "Fighting the fog of war: Automated incident detection for cloud systems," in *USENIX ATC*, 2021.
[18] Z. Li, N. Zhao *et al.*, "Actionable and interpretable fault localization for recurring failures in online service systems," in *ACM ESEC/FSE*, 2022.
[19] M. Momtazpour, J. Zhang *et al.*, "Analyzing Invariants in Cyber-physical systems using latent factor regression," in *ACM SIGKDD*, 2015.
[20] A. J. Oliner and A. Aiken, "Online detection of multi-component interactions in production systems," in *IEEE/IFIP DSN*, 2011, pp. 49–60.
[21] A. J. Oliner, A. V. Kulkarni, and A. Aiken, "Using correlated surprise to infer shared influence," in *IEEE/IFIP DSN*, 2010, pp. 191–200.
[22] A. B. Sharma *et al.*, "Fault detection and localization in distributed systems using invariant relationships," in *IEEE/IFIP DSN*, 2013.
[23] M. B. Sullivan, N. R. Saxena, M. O'Connor, D. Lee *et al.*, "Characterizing and mitigating soft errors in GPU DRAM," *IEEE Micro*, 2022.
[24] C. Tennant, A. Carpenter, T. Powers, A. Shabalina Solopova *et al.*, "Superconducting radio-frequency cavity fault classification using machine learning at Jefferson Laboratory," *Phys. Rev. Accel. Beams*, 2020.
[25] O. Tuncer, E. Ates, Y. Zhang *et al.*, "Online diagnosis of performance variation in HPC systems using machine learning," *IEEE TPDS*, 2019.
[26] Y. Wang, G. Li *et al.*, "Fast outage analysis of large-scale production clouds with service correlation mining," in *IEEE/ACM ICSE*, 2021.
[27] M. West and J. Harrison, *Bayesian forecasting and dynamic models*. Springer Science & Business Media, 2006.
[28] C. Wu, N. Zhao, L. Wang, X. Yang *et al.*, "Identifying root-cause metrics for incident diagnosis in online service systems," in *IEEE ISSRE*, 2021.
[29] Y. Zhao, D. Duplyakin, R. Ricci, and A. Uta, "Cloud performance variability prediction," in *ACM/SPEC ICPE*, 2021, pp. 35–40.