Causality Based Instant Root Cause Analysis for Microservices Failure

A technical paper prepared for presentation at SCTE TechExpo24

Mohamed Sharafath M
Engineer 3 – Machine Learning
Comcast India Engineering Center, Chennai, India
mohamedsharafath_mohamedimthiyas@comcast.com

Praveen Manoharan
Engineer 4 – Machine Learning
Comcast India Engineering Center, Chennai, India
praveen_manoharan@comcast.com

Aravindakumar Venugopalan
Director 1 – Machine Learning
Comcast India Engineering Center, Chennai, India
aravindakumar_venugopalan@cable.comcast.com

Table of Contents

List of Figures

1. Introduction

In modern distributed systems, the complexity and scale of operations often lead to challenging issues in identifying the root causes of system failures [1]]. Traditional ways of finding out why something happened might not work well with these complicated systems, especially if they only use metrics or logs data. The huge volume of data makes manual tracing and debugging of issues impractical in a time crunch situation. The inherent limitations of isolated data sources often result in prolonged downtime, increased operational costs, and hindered system performance.

Our proposed solution seeks to automate the construction of microservice dependencies by leveraging causal discovery techniques with multi-variate time-series data. With an increasing focus on explain-ability in many domains, causal inference has attracted much attention in the industry [2]]. In this paper, we consider a fault in microservices as an intervention in causal inference. The Bayesian-based causal inference algorithms [3] are applied to the constructed dependency graph tree at each level. This facilitates the swift identification of the likely root cause path of microservice failures. Such prompt analysis empowers site reliability engineers (SREs) to make informed, data-driven decisions. In this paper, we discuss how implementing Causality based instant Root Cause Analysis (RCA) methods in AI for Information Technology Operations (AIOps) platforms improves reliability for efficient triaging to reduce Mean Time to Repair (MTTR).

2. The Life Cycle of Instant RCA

Information Technology (IT) operations require constant monitoring of IT infrastructure, applications, and services to find and fix problems early. This includes monitoring network performance, server health, app availability, and other critical metrics. Regular maintenance activities such as patch management, upgrades, and backups are also parts of maintenance activities.

The following sections talk about the different stages of automated RCA in IT operations for microservices.

2.1. Data Collection

The data collection module is a crucial component in the RCA life cycle. This module is responsible for gathering diverse and comprehensive data from various sources, ensuring that the system has the necessary information to accurately identify, diagnose, and resolve issues within an IT infrastructure.

The key components in data collection are as follows:

2.1.1. Metrics and Monitoring Tools:

Gather metrics from monitoring tools that track central processing unit (CPU) usage, memory consumption, disk input/output (I/O), network traffic, and application performance [4] .

2.1.2. Configuration Management Databases:

Integrate with configuration management databases (CMDBs) to obtain information about the configuration and relationships of various monitoring metrics.

2.2. Structural Causal Model (Dependency Mapping)

The structural causal model (SCM) [5] is a powerful tool used in causal inference to model and understand the relationships and dependencies among variables in a system. When applied to dependency

mapping in IT or complex systems, SCMs provide a structured framework for comprehending and visualizing how changes or events in one part of the system can affect others.

Variables can represent various components such as servers, applications, network devices, databases, and user interactions. Each variable is associated with attributes like performance metrics, latency, throughput, memory usage, etc. To make it relatively straightforward, we handle variables that are time-series data. SCMs use directed acyclic graphs (DAGs) to represent causal relationships among variables. Nodes in the graph represent variables, while directed edges between nodes indicate causal influences.

In reality, there is never perfect information about the SCM that underlies data. Instead, there typically is only a bunch of observations about the underlying system. Causal discovery methods can be employed to identify causal relationships. One such example is the Peter Spirtes - Clark Glymour Momentary Conditional Independence (PCMCI) algorithm [6] to identify the appropriate causal links.

The PCMCI assumes that there are no instantaneous causal links between the variables. This is a reasonable assumption to make in the time-series setting of microservices monitoring.

A schematic representation of this process is described in the following Figure 1.

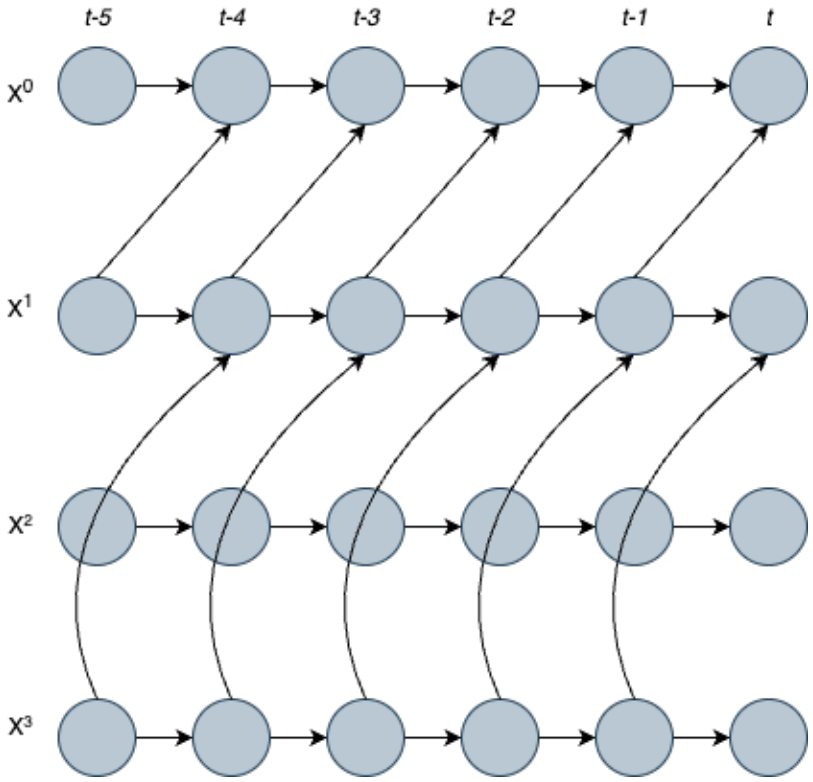

Figure 1 – Causal Dependency Identification

In this Figure 21, it is evident that PCMCI algorithms find causal dependencies between X^0 and X^1, X^1 and X^3 with 1-time lag. So, the final SCM constructed is - X^3 causes X^1 and X^1 causes X^0.

2.3. Root Cause Analysis using Causal Intervention

RCA plays a crucial role in reducing Mean Time to Detect (MTTD) and MTTR. RCA algorithms facilitate the rapid identification of outage origins, contrasting with manual operator efforts that involve scrutinizing multiple dashboards to isolate issues.

In the context of RCA, the utilization of a dependency graph depicting services and applications aids RCA algorithms in efficiently navigating and pin-pointing the accurate cause of incidents. The causal dependency graphs constructed using causal discovery methods are further evaluated and edited with the help of SRE team using their domain expertise around the system.

The RCA module which we introduce in this work, assumes that the faulty period is the intervention period. It integrates a structural Bayesian time-series model to evaluate how the time-series response metric might have evolved if the intervention had not occurred. In simpler terms, the model uses the pre-intervention period's multivariate time-series as a control to explain the outcome time-series at each level of causal dependency graphs. This approach helps identify probable anomalous nodes and ultimately determine the anomaly propagation path.

To facilitate causal inference from the SCMs, the metrics that share a common cause with a particular metrics are grouped together. The following Figure 2 explains the model building in more detail.

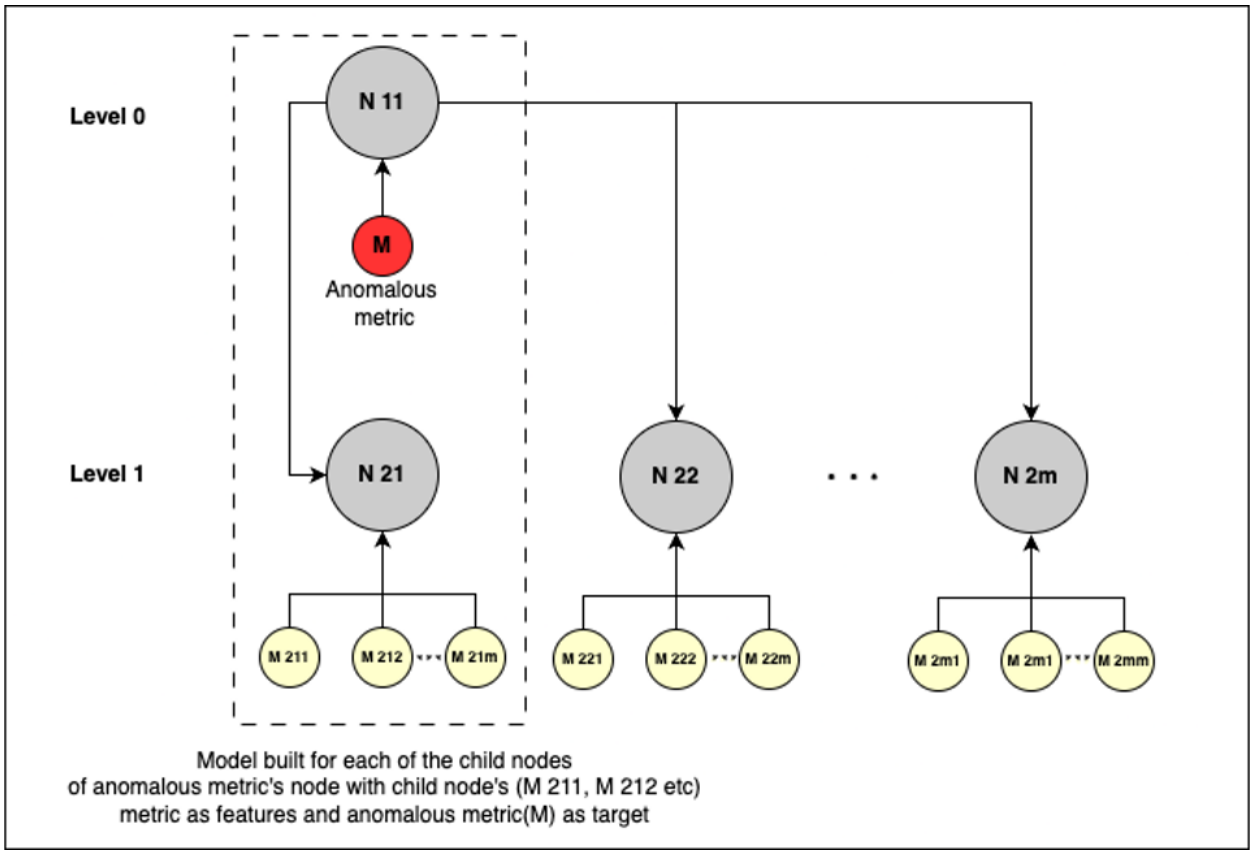

Figure 2 – Causal Intervention-Based Model Building

3. Instant RCA in AI For IT Operations Platform

The increasing push for digital transformation in organizations and the dynamism of cloud computing are presenting IT operations with obstacles that traditional management paradigms cannot handle [7] [8]]. AIOps is a promising technology that can mitigate the increasing complexity of IT management by utilizing AI and Big Data [9]]. AIOps platforms are defined as highly scalable software systems that ingest data from a variety of sources to perform comprehensive analyses. They enable stakeholders to identify patterns that can be used to analyze and identify the root cause of incidents [10] [11] .

The anomaly detection module in AIOps is a critical component that automates the identification of deviations or abnormalities in data patterns across IT systems. It collects and pre-processes data from diverse sources, selects appropriate anomaly detection algorithms, and trains models using historical data to establish normal behavior baselines. In real-time, these models continuously monitor incoming data, flagging and categorizing anomalies based on severity and impact. The module generates alerts and notifications for prompt response, supports RCA by pin-pointing underlying issues, and incorporates a feedback loop for model refinement. Through visualization and reporting tools, it provides actionable insights to improve system reliability, minimize downtime, and optimize operational performance, driving proactive IT management strategies.

3.1. Need for Instant RCA in AIOps

The conventional RCA module within AIOps systems relies on correlating actual time-series data or determining root causes based on correlations of time-series anomaly predictions using a dependency graph of services and applications. However, this approach faces scalability and cost challenges as it may necessitate building anomaly detection models for every metric under consideration, which will be of very high volume for most real-world applications. Consequently, there arises a pressing need for instant root cause analysis (Instant RCA) capabilities within AIOps frameworks. Instant RCA refers to the ability to pin-point the root causes of anomalies or issues swiftly and accurately in real-time, without the need for pre-built anomaly detection models for every metric. This capability streamlines the RCA process, enhances scalability, and enables proactive and efficient problem resolution within complex IT environments.

3.2. High-Level View of Instant RCA

A schematic representation of high-level view of Instant RCA is depicted in Figure 3.

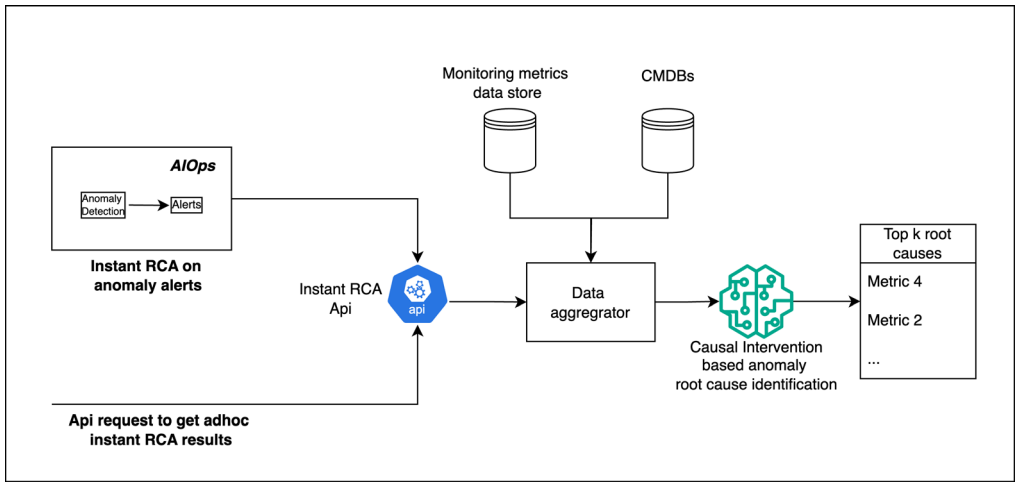

Figure 3 – High-Level View of Instant RCA

4. AIOps Instant RCA Case Study

Comcast's Development and Operations (DevOps) team manages significant complexity associated with various system metrics originating from multiple sources. Their primary challenge lies in performing RCA, particularly under time constraints during critical situations when issues directly impact customers. The team has benefited from the implementation of Instant RCA.

4.1. Triggering Instant RCA

The Instant RCA feature has been integrated into a channel primarily utilized by SREs for discussions related to metrics outages and monitoring.

4.2. Instant RCA Findings

Upon selecting the necessary metrics within a graphical user interface (GUI), specific anomaly timestamp or ad hoc timestamp, time-series granularity, aggregation type, and approximate anomaly duration, causal intervention-based regression models are constructed in the backend. These models are developed for each level in the dependency graph of nodes (fetched from CMDBs) to identify the paths through which anomalies propagate.

Results derived from the Instant RCA feature may include a depiction of the root cause traversal path which may depict nodes representing entities that are labeled to map to metrics monitored in AIOps. This way, RCA can traverse across multiple levels.

Traditionally, SREs have needed to manually examine each monitoring metric and traverse through dependent metrics to determine the probable root cause propagation path. However, the AIOps feature of Instant RCA leverages causality to deliver results within minutes. This significantly reduces the need for tedious manual work, allowing SREs to focus their valuable time on other critical aspects of their operations. Additionally, this feature greatly helps in reducing the MTTR, enhancing overall operational efficiency.

5. Applicability of Causal RCA to Telecom Network Device Outages

Network infrastructure troubleshooting is a multi-layered process, progressing from the initial identification of a general issue to the detailed RCA of a specific problem. Given the demonstrated success of causality-based RCA in handling time-series microservices data, this approach is highly applicable to the domain of network devices, where device statuses and outages are discrete in nature.

The process involves causal discovery to identify dependencies among network devices and determine the root cause of incidents. This method is particularly effective for the following reasons:

Discrete Nature of Data: Network device statuses and outages typically present as discrete events rather than continuous time-series data. Causality-based RCA can handle such discrete data effectively, allowing for accurate identification of the relationships between different network components.

Causal Discovery: By utilizing advanced causal discovery algorithms, it is possible to map out the dependencies among network devices. This involves analyzing various metrics and logs to uncover how different devices and components influence one another.

Root Cause Identification: Once the dependencies are established, causal inference techniques can be employed to pin-point the root cause of any observed incident. This involves simulating interventions and

analyzing the resultant changes in network performance, thereby identifying the specific device or interaction responsible for the issue.

The implementation of causality-based RCA in the network domain is not only feasible but also highly advantageous. It allows for a structured and systematic approach to troubleshooting, transforming vague problem identification into precise root cause determination.

6. Conclusion

The Instant RCA module in an AIOps platform is a pivotal component designed to enhance the operational efficiency and reliability of IT systems. By leveraging advanced machine learning algorithms, real-time data processing, and comprehensive analytical techniques, this module facilitates the swift identification and resolution of issues, minimizing downtime and maintaining service continuity. The Instant RCA module's integration with various data sources, such as performance metrics, and configuration management databases, ensures a holistic view of the IT environment.

Moreover, the module's ability to instantaneous root cause findings using causal intervention-based ML models significantly reduces the operational burden on IT teams, allowing them to focus on strategic initiatives rather than routine troubleshooting.

Contrary to traditional practices employed by SREs for RCA, the Instant RCA provides essential information to SREs promptly, facilitating more efficient troubleshooting of outages. casein one beta test, for example, it was estimated that the MTTD decreased significantly, from an average of 30 minutes to approximately 1 to 2 minutes.

In summary, the Instant RCA module is an indispensable tool in the AIOps platform, driving operational excellence through precise, automated RCA. Its implementation is essential for modern IT operations aiming to achieve higher efficiency, reduced downtime, and enhanced service reliability.

7. Acknowledgement

We would like to acknowledge our Comcast leaders in the USA: Rick Rioboli, Jan Neumann, Faisal Ishtiaq, Jim Cahill and Nawar Elmolla, and in India: Kannan Subramaniam and Harish Jayesh, for their support in the initiative. We would also like to acknowledge our colleagues at AI Technologies team: Nagaraj Sundaramahalingam, Nilesh Nayan, Aaditya Sharma, Jaswanth Duthaluru, Mahesh Yadav and Shivcharan Thirunavukkarasu for their contributions to the Research and Development (R&D) of our AIOps platform. In addition, we are grateful for the collaborative efforts and support extended by Nilesh Singh and his SRE team for their guidance as well as in validating and providing feedback on our algorithm performance. We also thank Kolammal Sankaranarayan for reviewing our paper and Nicholas Pinckernell for volunteering to present our work at SCTE TechExpo’24 on behalf of us.

Abbreviations

AI	artificial intelligence
AIOps	artificial intelligence for information technology operations
CMDB	configuration management database
CPU	central processing unit
DAG	directed acyclic graph
DevOps	development and operations
GUI	graphical user interface
I/O	input/output
IT	information technology
KPI	key performance indicator
MTTD	mean time to detect
MTTR	mean time to recovery
PCMCI	Peter Spirtes - Clark Glymour momentary conditional independence
RCA	root cause analysis
R&D	research and development
SCM	structural causal model
SCTE	Society of Cable Telecommunications Engineers
SRE	service reliability engineer
USA	United States of America

Bibliography & References

[1] Li Wu, Johan Tordsson, Erik Elmroth, Odej Kao *MicroRCA: Root Cause Localization of Performance Issues in Microservices* https://ieeexplore.ieee.org/document/9110353

[2] Mingjie Li, Zeyan Li, Kanglin Yin, Xiaohui Nie, Wenchi Zhang, Kaixin Sui, Dan Pei. 2022. *Causal Inference-Based Root Cause Analysis for Online Service Systems with Intervention Recognition* https://arxiv.org/abs/2206.05871

[3] https://www.sciencedirect.com/topics/social-sciences/causal-inference#:~:text=Introduction%3A%20Causal%20Inference%20as%20a,causal%20conclusions%20based%20on%20data.

[4] *What is APM (Application Performance Monitoring)?* https://aws.amazon.com/what-is/application-performance-monitoring/#:~:text=Application%20performance%20monitoring%20(APM)%20is,receive%20a%20positive%20application%20experience.

[5] Sarthak Chakraborty, Shaddy Garg, Shubham Agarwal, Ayush Chauhan, Shiv Kumar Saini. 2023. *CausIL: Causal Graph for Instance Level Microservice Data.* https://arxiv.org/abs/2303.00554

[6] Jakob Runge, et. al. 2017. *Detecting causal associations in large nonlinear time series datasets* https://arxiv.org/pdf/1702.07007

[7] Masood, A., & Hashmi, A. 2019. *AIOps: Predictive Analytics & Machine Learning in Operations*. Cognitive Computing Recipes, 359–382. https://doi.org/10.1007/978-1-4842-4106-6_7

[8] Levin, A., Garion, S., Kolodner, E. K., Lorenz, D. H., Barabash, K., Kugler, M., & McShane, N. 2019. *AIOps for a Cloud Object Storage Service*. 2019 IEEE International Congress on Big Data (BigDataCongress). https://doi.org/10.1109/bigdatacongress.2019.00036

[9] Paradkar, S. (2020). *APM to AIOps - Core Transformation. Global Journal of Enterprise Information System*. https://doi.org/10.18311/gjeis/2020

[10] Hongcheng Wang, Praveen Manoharan, Nilesh Nayan, Aravindakumar Venugopalan, Abhijeet Mulye, Tianwen Chen, and Mateja Putic. *AI for IT operations (AIOps) – Using AI/ML for improving IT Operations*. SCTE Fall Technical Forum Proceedings NCTA Technical Papers, 2022 https://www.nctatechnicalpapers.com/Paper/2022/FTF22_AIML02_Wang_3756

[11] Praveen Manoharan, Nilesh Nayan, Aaditya Sharma, Aravindakumar Venugopalan. *Building a scalable real-time ML inference platform for AIOps*. Volume-4 ISSUE-1, Lattice, The Machine Learning Journal by Association of Data Scientists, January 3, 2023 https://adasci.org/building-a-scalable-real-time-ml-inference-platform-for-aiops/

