Data Corruption Solutions

Bhoomika.S, Uma Devi Ramamoorthi

[1] *Student, School of science and computer studies, CMR University, Karnataka, India*
[2] *Associate Professor, School of science and computer studies, CMR University, Karnataka, India*

ABSTRACT

This collection of articles discusses various approaches to addressing data corruption, memory safety, and reliability challenges in different fields such as computing, network systems, and data analysis. One paper proposes a machine learning model in Python to overcome data corruption, achieving high accuracy rates in data recovery and protection. Another article introduces an aggregated services approach to review network performance and prevent data corruption during data transmission. There’s a focus on mitigating silent data corruptions (SDCs) through detailed analysis of microarchitectural structures in modern processors and proposing tailored memory safety techniques. A paper discusses a solution called meta-Safer, which verifies metadata in Web Assembly (Wasm) memory to prevent attacks and ensure memory safety. Furthermore, there are discussions on software solutions for soft error protection in microprocessors and the development of more general schemes to detect and prevent data corruption. Overall, these articles offer insights into innovative methods and technologies aimed at enhancing data protection, recovery, and reliability in various domains.

Keyword: - Silent Data Corruption (SDC). Memory Safety, Machine Learning, Data Recovery, Network Performance.

1. INTRODUCTION

The advancements in technology have led to a significant focus on combating data corruption through innovative approaches, particularly utilizing machine learning models in Python like LOF and K-means clustering. These models have been instrumental in fine-tuning methods aimed at recovering corrupted data and preventing future corruption. Additionally, there's a growing concern regarding silent data corruptions (SDCs) in various sectors, including academia, manufacturing, and hyperscale operations. Researchers are addressing this issue by analysing critical microarchitectural structures in modern processors prone to SDCs. Moreover, innovative tools like Spots DC have been introduced to visualize and understand how programs handle SDCs, aiding in improving program resilience. Other research areas include protecting data integrity in distributed networks and devising methods to ensure memory safety in technologies like Web Assembly (Wasm). Furthermore, addressing soft errors in microprocessors through software solutions, such as the proposed general scheme with coarse-grained scheduling and asymmetric control-flow signatures, shows promising results in reducing silent data corruptions and enhancing program efficiency.

2. LITERATURE SURVEY

Charvi Bannur et.al with the quest of technology developed working model using machine learning in python language ,which will help to overcome the data corruption in the data field which is playing a major role to get the data back from the corrupted data and develop the methods which are efficient to get back the corrupted data and prevent them from corruption in the future ,they are mainly focusing on the corrupted data rather than clusters, with the help of the python models like LOF(Local Outlier Factor),K-means clustering and other models helped in fine tuning the model which will help to attain the high accuracy and recall of the model which will help to develop in providing high accuracy in the field of data corruption , data protection and data recovery in all the fields so, that the problem of loss of data can be overcome with the help of this model they attained a with an accuracy of 96.35% for clustered data and 99.04% for linear data. in an effective manner.

George Papadimitriou et.al. the rising concern of silent data corruptions (SDCs) in academia, manufacturing, and hyperscale operations. Recent studies by Meta and Google highlight a surprising increase in SDC incidents linked to modern microprocessors, particularly in large data centres. Despite its severity, there's a lack of detailed analysis on which microarchitectural locations within complex processors are prone to SDCs. The paper presents an in-depth examination of various critical microarchitectural structures in modern out-of-order microprocessors causing SDCs. Key observations include the varying impact of different hardware structures, the influence of instruction-related parameters, the role of operating systems, and vulnerable byte positions in data words. These findings provide valuable insights for devising hardware and software solutions to mitigate the risk of silent data corruption.

Junchi Ma er.al they talk about the issue of soft errors in electronic devices caused by radiation as technology gets smaller. Silent data corruption (SDC) is a big concern, where errors occur without any signs. They propose a metric called Output Vulnerability Factor (OVF) to figure out which parts of a program are most likely to cause SDCs. This helps prioritize which parts to protect. They use a model called Enhanced Dynamic Dependence Graph (eDDG) to calculate OVF and identify where errors might spread. By focusing on the most vulnerable parts, their method detects SDCs 65% of the time, which is better than previous methods. It's a step forward in making electronic devices more reliable.

Kaustav Chatterjee et.al. They have discussed a method for organizing PMU signals to ensure that data can be recovered even if some of it is corrupted. They develop analytical relationships to understand how dense the space covered by a set of measurements is. By grouping signals based on their phase angles and amplitudes, they can minimize errors and improve the chances of recovering accurate data. This approach relies on a technique called Robust Principal Component Analysis. They test their method on both simulated and real PMU data from a utility company in the US.

Kaveri Mahapatra et.al. They introduce a way to protect phasor measurement unit (PMU) data from malicious tampering, especially for wide-area damping control. They treat the problem of spotting corrupted PMU data like solving a puzzle and use a special algorithm to fix the tampered signals. They compare their method with another technique to see which one works better. They show how their approach can handle different types of attacks on the PMU data, like missing data. Finally, they demonstrate the effectiveness of their method in improving the stability of a power system in a specific region with multiple machines.

Marcel Klaes et.al proposed an aggregated services in order to review the performance of the data distribution on the network performance, where in the current generation the data is been transmitted from one system to another during this process lot of corruption in the data like data loss, data corrupted, unwanted data transmission might take place to avoid this in the distributed network where the main reason for this was the power of authority in the data handling process was been transmitted from one level to another level ,where people who doesn’t have the knowledge of accessing or protecting the data where given the authority in this process data corruption used to take place .To control this in general there was lot of strategies which was undergone and one of them was aggregated service state description which gave a brief description about the errors which used to take place during the process with the help of it they understood the process of the results by comparing the designs and setups of the services.

Moona Yakhchi et.al. They talk about a problem called silent data corruptions (SDCs), usually caused by radiation. Traditional fixes for SDCs are expensive and use a lot of resources. Researchers are now looking for cheaper and better ways to deal with SDCs. Finding which parts of a program are prone to SDCs is hard and usually requires time-consuming tests. The article suggests a new way to find and reduce SDCs in whole programs without needing to do these tests. They use a mix of machine learning and a special algorithm to predict which parts of the program might have SDCs. Tests show this method is 99% accurate and only slows down performance by 58%.

Moslem Didehban et.al They have discussed a common problem with microprocessors called soft errors, where transient faults can cause issues. One way to protect against these errors is through software solutions. For example, the nZDC scheme can detect errors by checking memory writes and using a control-flow mechanism. However, this mechanism is specific to certain architectures and has some weaknesses. This paper suggests replacing it with a more general scheme and introduces two new techniques, coarse-grained scheduling and asymmetric control-flow signatures, to better detect tricky control-flow errors. Tests on different hardware components show that this new approach reduces silent data corruptions by 85% compared to nZDC and even makes programs run about 37% faster.

Na Yang et.al spoke about silent data corruption (SDC), which is a really bad kind of error that can happen in computer programs. They want to find out which instructions in a program are likely to cause these errors, so they can protect those parts better. But current methods for finding these instructions need a lot of testing and don't work well for different programs or inputs. So, they came up with PVInsiden, a new way to find these vulnerable instructions using machine learning. They test it using only 35% of the usual testing, and it still identifies 85% of the vulnerable instructions with 80% accuracy, which is pretty good. PVInsiden works well for different kinds of programs and inputs too.

Radu Ioan Ciobanu et.al. They spoke about a new approach to computing called drop computing, which tries to solve problems with traditional mobile cloud computing, especially with the rise of the Internet of Things and many connected devices. In drop computing, devices can share data and tasks with each other, the cloud, or nearby devices using social connections. The paper focuses on how devices can exchange data directly with each other using close-range communication, but there can be problems with data getting corrupted or being tampered with. They suggest different ways to make sure the data stays correct, like using a rating system or checking the data carefully. Their experiments show that their methods work well and can keep data safe in the network, sometimes reaching 100% accuracy.

Suhyeon Song et.al They spoke about Web Assembly (Wasm), a technology used to run native code efficiently in web browsers. It has become popular for creating lightweight web services that are fast and use less data. However, because of optimizations made to improve performance, there are concerns about memory safety. Wasm's memory structure, called linear memory, can be vulnerable to attacks, particularly through manipulation of metadata. Attackers can exploit these vulnerabilities to modify data or execute malicious code. While solutions for memory safety exist for other languages and architectures, they may not directly apply to Wasm. Therefore, there's a need for new techniques to ensure memory safety in Wasm. The paper proposes a solution called meta-Safer, which verifies metadata in Wasm memory to prevent attacks.

Venkat Ram Subramanian et.al they explained how to understand complex networks by modelling them as graphs. In these networks, nodes represent different parts, and links show how they interact. The main goal is to figure out how one-part influences another over time. However, data from these systems can be messy due to things like time differences, lost packets, and noise. The article explains that trying to understand these systems with messy data can lead to wrong conclusions. They develop a way to check if the data is reliable enough to draw conclusions from. Their method works for systems that are not straightforward and have feedback loops. They test their method and show that it gives reliable results most of the time.

Venkat Ram Subramanian et.al. they discuss how we can understand complex systems by representing them as graphs, with nodes as different parts and links showing how they connect. Previous research has shown that for certain types of systems, we can figure out their structure by looking at the data they produce. However, real-world data is often messy, with errors like wrong time-stamps and noise. The paper shows that trying to understand these systems with messy data can lead to wrong conclusions, particularly in terms of the connections between different parts. They explain that errors in the data affect nearby connections more, and this applies not just to certain types of systems but also to others like Markov random fields.

Yu-Shun Hsiao et.al. They focus on making sure self-driving vehicles are safe and resilient. They introduce ROSFI, a method to check how well a robot system handles silent data corruption (SDC), where data errors aren't immediately obvious. They use drones as an example and find that certain tasks like planning and control are more affected by SDCs than others. They offer two ways to detect and recover from these errors without using too much extra resources. They test their methods in different virtual environments and show they work well. Their method is open-source, so others can use it to make sure their own robot systems are robust too.

Zimin Li et.al discussed how as technology advances, high-performance computing systems become more prone to errors, like random bit flips. Some errors may not affect much, while others can cause serious issues, known as silent data corruption (SDC). Current methods use random bit flips to test program resilience, but summarizing results is hard. So, they introduce Spots DC, a visualization tool to help understand how a program handles SDCs. Spots DC shows where errors occur in the code, which bits are affected, and when they happen during execution. This tool also helps study how well the code is protected and provides insights to improve fault injection tests. Overall, Spots DC aims to make analysing and improving program resilience easier.

3. PROPOSED METHOD

3.1 Machine Learning Models:

- LOF (Local Outlier Factor)
- K-means clustering
- Special algorithms for predicting silent data corruptions (SDCs)

3.2 Data Analysis and Visualization:

- Visualization tools like Spots DC for understanding program behavior regarding SDCs
- Analyzing complex networks through graph modeling

3.3 Algorithmic Approaches:

- Output Vulnerability Factor (OVF) for identifying SDC-prone parts of programs
- Enhanced Dynamic Dependence Graph (eDDG) for calculating OVF and detecting SDCs
- Robust Principal Component Analysis for organizing PMU signals and minimizing errors
- Special algorithms for detecting and fixing corrupted PMU data
- Coarse-grained scheduling and asymmetric control-flow signatures for detecting soft errors in microprocessors
- Machine learning-based approach (PVInsiden) for identifying vulnerable instructions

3.4 Data Integrity and Security:

- Protection methods for PMU data against tampering
- Memory safety techniques like meta-Safer for Web Assembly (Wasm) memory verification

3.5 System Resilience and Testing:

- ROSFI method for detecting and recovering from SDCs in robot systems
- Spots DC for analyzing and improving program resilience
- Aggregated service state description for reviewing data distribution performance and error analysis in distributed networks.

4. EXPERIMENTAL SETUP

- Developed a machine learning model using Python to overcome data corruption.
- Focused on corrupted data rather than clusters, using techniques like LOF and K-means.
- Achieved high accuracy in data recovery and prevention.
- Examined microarchitectural structures in modern processors causing silent data corruptions (SDCs).
- Provided insights into vulnerable hardware components and software solutions.
- Addressed soft errors in electronic devices caused by radiation.
- Introduced a metric called Output Vulnerability Factor (OVF) to prioritize protection of vulnerable program parts.

- Used Enhanced Dynamic Dependence Graph (eDDG) for error detection.
- Proposed a method for organizing PMU signals to ensure data recovery even with corruption.
- Utilized Robust Principal Component Analysis for minimizing errors in signal grouping.
- Introduced a method to protect PMU data from tampering, especially for wide-area damping control.
- Developed an algorithm to spot and fix corrupted PMU data, improving power system stability.
- Proposed aggregated services to review data distribution performance and prevent corruption.
- Addressed issues in distributed networks, particularly related to authority levels and data handling.
- Suggested a machine learning-based approach to predict and reduce silent data corruptions in programs.
- Achieved high accuracy with minimal performance impact.
- Discussed software solutions to mitigate soft errors in microprocessors.
- Introduced new techniques for error detection and reduction, improving program performance.
- Developed PVInsiden, a machine learning-based method to identify vulnerable instructions in programs.
- Achieved high accuracy with reduced testing requirements.
- Explored drop computing and its challenges in ensuring data integrity.
- Proposed methods for data verification and protection in drop computing environments.
- Addressed memory safety concerns in Web Assembly (Wasm) using metadata verification.
- Proposed meta-Safer as a solution to ensure memory safety in Wasm.
- Developed methods for understanding complex networks and ensuring data reliability.
- Proposed techniques for checking data reliability and drawing accurate conclusions.
- Introduced ROSFI for checking self-driving vehicle resilience against silent data corruptions.
- Provided methods for detecting and recovering from errors without excessive resource usage.
- Developed Spots DC, a visualization tool for understanding program resilience against silent data corruptions.
- Facilitates analysis of error occurrences and protection effectiveness.

5. CONCLUSIONS

Researchers are addressing data corruption issues across diverse fields, employing innovative techniques to enhance data integrity and reliability. They've developed Python-based machine learning models and algorithms to efficiently repair and prevent data corruption, achieving high accuracy in data recovery. Insights into silent data corruptions (SDCs) in modern microprocessors have led to strategies to mitigate risks associated with hardware errors. Metrics like Output Vulnerability Factor (OVF) prioritize protection for vulnerable parts of computer programs. Techniques to detect and reduce soft errors in microprocessors have improved program reliability and performance. Methods such as ROSFI ensure the resilience of robot systems against SDCs, enhancing safety in autonomous vehicles. Algorithms and strategies safeguard critical infrastructure data from tampering and ensure data integrity in distributed networks. Visualization tools like Spots DC aid in analysing and improving program resilience, while techniques to verify data reliability in complex networks enable more accurate decision-making. These efforts collectively advance data security, fostering reliability and security across various domains, from microprocessor design to network communication and autonomous systems.

6. REFERENCES

[1]. Charvi Bannur, Chaitra Bhat, Kushagra Singh, Shrirang Ambaji Kulkarni, Mrityunjay Doddamani.,
” Comprehensive Data Corruption Detection Algorithm”, IEEE Access (Volume: 11), 06 March 2023.
[2]. George Papadimitriou; Dimitris Gizopoulos,” Silent Data Corruptions: Microarchitectural Perspectives”, IEEE Transactions on Computers (Volume: 72),13 June 2023.
[3]. Junchi Ma, Zongtao Duan, Lei Tang,” A Methodology to Assess Output Vulnerability Factors for Detecting Silent Data Corruption”, IEEE Access (Volume: 7),22 August 2019

[4]. Kaustav Chatterjee; Nilanjan Ray Chaudhuri; George Stefopoulos,” Signal Selection for Oscillation Monitoring with Guarantees on Data Recovery Under Corruption”, IEEE Transactions on Power Systems (Volume: 35),07 May 2020.
[5]. Kaveri Mahapatra, Mahmoud Ashour, Nilanjan Ray Chaudhuri, Constantino M. Lagoa, “Malicious Corruption Resilience in PMU Data and Wide-Area Damping Control”, IEEE Transactions on Smart Grid (Volume: 11),12 July 2019.
[6]. Marcel Klaes, Jannik Zwart Scholten, Anand Narayan, Sebastian Lehnhoff, Christian Rehtanz, “Impact of ICT Latency, Data Loss and Data Corruption on Active Distribution Network Control”, IEEE Access (Volume: 11), 08 February 2023.
[7]. Moona Yakhchi; Mahdi Fazeli; Seyyed Amir Asghari,” Silent Data Corruption Estimation and Mitigation Without Fault Injection.”, IEEE Canadian Journal of Electrical and Computer Engineering (Volume: 45),07 September 2022.
[8]. Moslem Didehban; Hwisoo So; Prudhvi Gali; Aviral Shrivastava; Kyoung woo Lee, “Generic Soft Error Data and Control Flow Error Detection by Instruction Duplication”, IEEE Transactions on Dependable and Secure Computing (Volume: 21), 16 February 2023.
[9]. Na Yang; Yun Wang,” Identify Silent Data Corruption Vulnerable Instructions Using SVM”, IEEE Access (Volume: 7),17 March 2019.
[10]. Radu Ioan Ciobanu; Vlăduț Constantin Tăbuşcă; Ciprian Dobre; Lidia Băjenaru; Constandinos X. Mavromoustakis,” Avoiding Data Corruption in Drop Computing Mobile Networks”, IEEE Transactions on Automatic Control (Volume: 66), 10 March 2019.
[11]. Suhyeon Song; Seonghwan Park; Donghyun Kwon,” meta-Safer: A Technique to Detect Heap Metadata Corruption in Web Assembly”, IEEE Access (Volume: 11), 26 October 2023.
[12]. Venkat Ram Subramanian; Andrew Lamperski; Murti V. Sala Paka,” Effects of Data Corruption on Network Identification Using Directed Information”, IEEE Transactions on Automatic Control (Volume: 67), 01 July 2021.
[13]. Venkat Ram Subramanian; Andrew Lamperski; Murti V. Sala Paka,”: Network Structure Identification from Corrupt Data Streams”, IEEE Transactions on Automatic Control (Volume: 66),26 November 2020.
[14]. Yu-Shun Hsiao, Zishen Wan, Tianyu Jia, Radhika Ghosal, Abdulrahman Mahmoud, Arijit Ray Chowdhury,” Silent Data Corruption in Robot Operating System: A Case for End-to-End System-Level Fault Analysis Using Autonomous UAVs”, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (Volume: 43), 13 November 2023
[15]. Zimin Li, Harshitha Menon, Dan Maljovec, Yarden Livnat, Shusen Liu, Kathryn Mohror, Peer-Timo Bremer,” Spots DC: Revealing the Silent Data Corruption Propagation in High-Performance Computing Systems”, IEEE Transactions on Visualization and Computer Graphics (Volume: 27),15 May 2020.