

A Survey of Fault Tolerance Technique in Distributed System

Rand Alamleh and Nammer El Emam

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

November 14, 2024

A Survey of Fault Tolerance Techniques inDistributed Systems

Rand Alamleh^{1,a} and Nammer N. El. Emam^{1,b}

¹Computer Science, College of Computer Sciences and Informatics, Amman Arab University a)Corresponding author: 202230132@aau.edu.jo b)n.emam@aau.edu.jo

Abstract. This paper presents a thorough survey of fault tolerance mechanisms in distributed systems. It examines potential failure factors, available mechanisms, and their foundations, focusing on mechanisms explicitly developed for distributed systems. This paper summarizes how fault-tolerance techniques can be combined to provide various dependability characteristics. The primary goal of this paper is to serve as a guide to the extensive research and development activity in the domain of distributed systems, examining the current fault tolerance mechanisms and highlighting future avenues for research aiming to help in identifying areas for further exploration and innovation, providing a roadmap for their future work and emphasizing the significance of their contributions to the research community.

Keywords: Fault Tolerance, Distributed Systems, Reactive Fault Tolerance, Proactive Fault Tolerance, Replication technique, Checkpointing technique.

INTRODUCTION

Distributed systems are crucial for modern applications and rely on fault tolerance mechanisms such as replication, high redundancy, and high availability to prevent service unavailability caused by system failures [1]. Due to their complex nature and the potential for various breakdowns, the significance of these mechanisms in ensuring the dependability and accessibility of these systems cannot be overstated. This paper delves into the specific mechanisms developed to reduce the impact of such breakdowns, enabling systems to maintain their functionality and efficiency. Fault tolerance in a distributed system refers to its ability to continue providing services to the user in the presence of faults or defects. Some faults are deviations of the system from the specifications of its components, while others are unexpected behavior exhibited by its components. Mechanisms, whether or not based on a theoretical foundation, are designed for practical use in all distributed systems to ensure dependability without prior knowledge of the distribution of the faults. These varying complexity mechanisms are not just theoretical concepts but are practical and applicable in real-world scenarios, instilling confidence in their potential use. Furthermore, fault-tolerance mechanisms in distributed query processing systems, like rollback-recovery mechanisms, enhance system resilience and performance, especially in long-running queries, by reducing the cost of fault tolerance provision and protocol overheads [2].

A fault can be categorized based on computing resources and time. A failure occurs during computation on system resources and can be classified as omission failure, timing failure, response failure, and crash failure. Failures can be categorized as follows:

- 1**. Permanent**: These failures occur due to accidentally cutting a wire or power breakdowns. They are easy to
	- reproduce and can cause significant disruptions. Some parts of the system may not function as desired. 2. **Intermittent**: These failures appear occasionally and are mostly overlooked during system testing. They occur when the system is in operation, making it hard to predict the extent of damage they can cause.
	- 3. **Transient**: These failures are caused by inherent faults in the system but can be corrected by retrying the system back to its previous state, such as restarting software or resending a message. They are widespread in computer systems.

RELATED WORKS

Over the past few years, some previous surveys have explored fault tolerance in distributed systems. We can summarize these efforts as follows. Ledmi, Bendjenna, and Hemam (2018) [3] pointed out in a survey the different techniques that tolerate faults in distributed systems in a detailed manner. They briefly discuss proactive and reactive faulttolerance approaches. Next, they focus mainly on the Check-pointing technique in distributed systems and highlight that many systems do not adequately address load balancing, availability, and reliability. Examines a range of fault tolerance mechanisms and tools, such as LA-MPI, Co-Check-MPI, and the Globus Toolkit, to guarantee overall reliability in distributed systems. Finally, they outline future research highlighting the need for a balanced approach between proactive and reactive fault tolerance techniques. Dhawan, Ahmad, and Tripathi (2022) [4] presented a systematic review discussing the concept of fault tolerance in distributed systems, highlighting faults as the root causes of errors and system failures. They emphasized the importance of fault detection and recovery techniques in real-time

distributed systems to prevent system failures, addressing various faults, including network, processor, and service expiration. Furthermore, the literature review in their study underscored the importance of fault tolerance mechanisms in distributed systems, showcasing the effectiveness of diverse approaches in mitigating system failures and ensuring continuous functionality. They further described the need to implement fault-handling approaches to maintain system integrity and continue operating seamlessly despite unexpected challenges. Future research directions were proposed, encompassing the amalgamation of graph and non-graph-based strategies for advancing high fault tolerance systems, thussignifying a sustained emphasis on augmenting fault tolerance capacities within distributed systems.

This review provides a more nuanced and updated perspective on fault tolerance techniques compared to earlier surveys. It incorporates recent trends such as MLBFT, blockchain-based checkpointing, ML load balancing, and quantum-resistant BFT, which are typically underrepresented in older literature.

FAULT TOLERANCE TECHNIQUES

Fault tolerance is a vital system property that maintains appropriate behavior following hardware and software faults [5]. Fault coverage is a significant parameter in fault-tolerant systems design and analysis. The fault coverage provided in a system can largely influence its reliability, safety, and other properties [1]. However, the system can now perform despite failure because of the fault-tolerant back-ups.

There are two main categories to classify fault tolerance techniques: reactive fault tolerance and proactive fault tolerance. Both approaches are essential for ensuring distributed systems' reliability, availability, and performance in the face of faults and failures.

Reactive Fault Tolerance

Reactive fault tolerance methodologies are based on the replication approach and concentrate on identifying and restoring faults after they manifest [11].

Replication

Having replication in distributed systems is crucial for maintaining fault tolerance. With replication, the system can duplicate and store its current state across several nodes, allowing it to recover and resume operation from the most recent replicated state in the event of a failure, as shown in Fig. 1. This is particularly important in fields like medicine, defense, aviation, telecommunication, and space, where system failure could have severe consequences [12]. Replication continuously duplicates and saves the system's state across multiple nodes, ensuring it can revert to a previously consistent state if a failure occurs. Critical applications can keep running without interruption, even with potential shortcomings [13]. The simple concept behind this replication methodology is to dock numerous replications of tasks to an already running task and submit them on the various hosts to avoid hanging all replicated tasks; the task itself would succeed in execution.

FIGURE 1. Replication technique

In their 2019 study, Smith and Kumar [14] delved into an advanced primary backup replication technique focused on optimizing communication protocols between primary and backup nodes to minimize latency. Their findings revealed that this method decreased latency by 30% compared to conventional mechanisms. However, they also noted that network partition scenarios may lead to database inconsistencies when the primary and backup nodes are isolated. Despite this drawback, the optimized method significantly reduced response time across symbionts during typical loads. In a separate study published in 2020, Wang and Zhang [15] introduced a novel adaptive replication technique that dynamically adjusts the number of replicas based on real-time workload and system conditions to balance performance with resource utilization. Notably, they found that frequent changes due to this method can result in significant overhead, particularly in rapidly changing operating environments. Nevertheless, these modifications yielded a substantial 25% improvement in data availability under workloads characterized by varying load levels and burstiness, highlighting the practicality of this approach in coping with unpredictable demand.

Checkpointing

Checkpointing is a critical feature in distributed systems, which is vital in ensuring fault tolerance. In the unfortunate event of a system failure, the checkpoint functionality allows for the preservation of ongoing distributed processes and associated data, thereby facilitating recovery and the resumption of operations from the last recorded checkpoint. This feature is significant in high-stakes industries such as medicine, defense, aviation, telecommunication, and space, where

system failures could potentially lead to catastrophic consequences. Figure 2 illustrates the various checkpointing techniques to safeguard the system's state. These techniques enable the system to revert to the saved state in the event of a failure, thereby ensuring that critical applications can continue uninterrupted without any significant impact.

FIGURE 2. Checkpointing technique

To reduce overhead, Patel and Gupta's (2019) [17] research focused on incremental checkpointing, which only records changes made since the last checkpoint. This approach effectively reduces checkpointing overhead by 45% without compromising system dependability, making it suitable for low-resource IoT systems even though it may miss transient states that cause discrepancies. Lee and Chen's (2020) article presents an adaptive checkpointing technique that modifies checkpoint intervals based on system performance and failure forecasting. Despite the high computational expense associated with failure prediction algorithms, this strategy enhances system operation by 20% through optimized checkpoint intervals, demonstrating its advantages in high-performance computing settings. According to Johnson and Brown (2021), cloud-based checkpoints ensure high availability by using cloud storage for checkpoint storage and improve recovery times by 30% after failures. Checkpointing is effective in distributed application environments despite network latency's possible effects on applications' performance. In case of an unsuccessful attempt, the recovery mechanism retrieves the transaction's last known checkpoint state from memory and carries out additional operations. [3].

Proactive Fault Tolerance

Proactive fault tolerance strategies strive to avert or alleviate faults before they emerge by anticipating and resolving potential challenges beforehand.

Load Balancing

In a distributed system, load balancing distributes the loads among processing nodes in Fig. 3 so that some have minimal work to do and others have none at all. Processes ought to be dispersed across processing nodes to optimize the efficiency of each process's completion and the efficient utilization of processing nodes [19].

Real-time monitoring is proposed in Ahmed and Wilson's (2019) research, which also applies a dynamic loadbalancing mechanism to distribute the resource among nodes. This method decreases query response time by 35%, demonstrating appropriate tooling for fixing distributed data responses under mixed loads, despite being unfeasible and possibly unstable during peak loads, reaching their goal of proving its efficiency in controlling the performance of distributed databases. Kim and Park (2020) propose an algorithm that utilizes a machine learning technique for load balancing, predicting future loads, and distributing them effectively between servers; while the approach may incur high computational costs for model training and inference, it achieves up to 25% higher load distribution efficiency compared to static methods, illustrating the potential benefits of machine learning in optimizing load balancing. An energy-aware load-balancing strategy that maximizes workload distribution and minimizes energy consumption is presented by Clark and Moore (2021). Using this method, energy consumption was reduced by 15% with little effect on system performance, demonstrating its efficacy in managing energy efficiency within data centers despite the tradeoff between energy savings and performance levels.

FIGURE 3. Load Balancing

Tolerance (BFT)

Distributed systems rely on Byzantine Fault Tolerance (BFT) in Fig. (4) to ensure continued functionality even if some components are faulty or behave maliciously. BFT allows the system to reach a consensus among its participants so that the correct nodes can agree on the system's state, even in the presence of malicious nodes [23]. In practical terms, BFT adds an extra layer of trust by allowing the system to agree even in the face of adversarial behavior, thereby improving the system's reliability and security. Achieving consensus in a BFT system involves complex agreement processes designed to ensure safety and liveness properties with high probability, even in the presence of faults or malicious actors.

FIGURE 3**.** Byzantine Fault Tolerance (BFT) [24]

Byzantine fault tolerance is a fundamental challenge in distributed systems research and is a relatively general problem. The primary goal of Byzantine fault tolerance research is to build practical and secure protocols. One way to practically achieve Byzantine fault tolerance is with secret sharing schemes, which work by hiding information from the leaders of the distributed system, who could be malicious, and distributing results to all system members in a verifiable way.

In Wang, Liu, and Xu (2020) [24], an (ABFT) protocol was developed which is used explicitly for cloud systems; the protocol works on modifying the traditional (BFT) approach to handle unusual system log activities, which causes an enhancement in fault tolerance and efficiency in changing environments, while on the other hand, it comes with extra computational and complexity costs also the paper shows that ABFT is suitable for large-scale cloud systems that often experience varying workloads indicating that it improves fault tolerance by 30% and system responsiveness by 25%.

Scalable Peer-to-Peer BFT (SPP-BFT) protocol presented by Zhang, Li, and Yang (2022) [25] for peer-to-peer distributed systems. This approach uses a decentralized consensus and reputation-based voting to improve scalability and fault tolerance regarding its strengths in scalability and resilience; SPP-BFT faces issues with reputation management and potential centralization in their study. It shows that SPP-BFT can manage up to 1000 peers, significantly lower message complexity, and improve fault tolerance by 25%.

Liu and Zhao's (2024) [26] research presented a BFT technique enhanced with machine learning to advance fault detection and mitigation in distributed systems. The ML-BFT approach integrates machine learning to predict faults and adjust the BFT protocol, strengthening the system's ability to handle changing fault patterns. Nonetheless, this integration adds to computational overhead. Their research demonstrates a 35% improvement in fault detection accuracy and a 20% boost in system resilience through machine learning and BFT techniques.

Combined Techniques

Researchers have recently explored integrated approaches combining fault tolerance techniques to improve system resilience and performance. These combined techniques aim to utilize the strengths of different methodologies. In this section, the paper will discuss most of the recent advancements in these combined fault tolerance techniques, emphasizing their contribution to improving system efficiency and reliability across a range of computing environments. Taylor and Evans (2022) [27] combined replication and checkpointing techniques to improve fault tolerance in distributed systems; combining these methods introduces higher coordination complexity but enhances the system resilience by 35%, Emphasizing the advantages of using them together to increase reliability. While Lin, Zhang, and Wu (2019) [28] studied how to improve reliability in distributed systems by combining Byzantine fault tolerance (BFT) and checkpointing, the Hybrid-BFT method uses periodic checkpoints to save system states and recover quickly from faults it shows that it is a Hybrid-BFT is a feasible solution for fault-tolerant distributed systems cause it leads to a decrease in recovery time by 40% and enhances system reliability by 30% but it faces challenges with maintaining frequent checkpoints due to overhead. A multi-layered strategy that distributes duties evenly among multiple layers and duplicates critical information is introduced by Patel and White (2023) [29] research paper; this approach improves data accessibility and duties handling by 40%, demonstrating its efficacy in a complex distributed environment. A combined approach that merges load balancing with dynamic replication, adjusting replication levels

and load distribution based on real-time system performance and fault occurrence to ensure efficient use and management of resources is presented in Li and Zhang (2020) [30] research which shows a 25% increase in system throughput and a 20% decrease in fault recovery time with this technique, benefiting distributed systems with variable workloads. An approach that merges (BFT), replication, and load balancing was presented by Wang and Zhao (2023) [31], which is referred to as Adaptive Fault Tolerance (AFT); the research shows that AFT magnify fault tolerance by 50% and uplift system performance by 30%, which position it as a resilient solution for dynamic distributed environments in their methodology they dynamically modifies replication and load-balancing strategies in response to real-time system conditions to augment fault tolerance and performance. Patel and Desai (2022) [32] presented a methodology that combines checkpointing with load balancing to deliver a scalable approach to fault tolerance. The proposed framework, called Checkpoint-Load Balance (CLB), employs checkpointing to regularly preserve system states and load balancing to allocate computational tasks across various nodes evenly. This technique improves scalability and fault recovery. The investigation reveals that CLB boosts system scalability by 30% and diminishes fault recovery duration by 25%, presenting a well-rounded remedy for fault-tolerant distributed systems.

DISCUSSION

The most recent developments in distributed systems' fault tolerance demonstrate a range of creative strategies. A synopsis of each research is given in Table 1, emphasizing the main ideas. Researchers focus on optimization and suggest adaptive replication for cloud systems to increase scalability and robustness. Checkpointing strategies are also advancing, focusing on adaptive checkpointing for high-performance systems and lightweight checkpointing methods for IoT environments. Regarding load balancing, some researchers utilize machine learning for resource distribution management, while others integrate load balancing with Byzantine fault tolerance for peer-to-peer systems. Hybrid approaches that combine different fault tolerance techniques are gaining popularity to address various system requirements. Moreover, a unified framework is being proposed to adaptively incorporate fault tolerance, replication, and load balancing, focusing on scalable checkpointing and load balancing to enhance system resilience and resource management. These strategies demonstrate the ongoing evolution in fault tolerance and load balancing, indicating a shift toward more integrated and adaptive solutions.

TABLE 1. Summary of literature review

CONCLUSION

The recent advancements in fault tolerance and load balancing in distributed systems demonstrate a profound understanding of evolving strategies and methodologies. Efficient replication techniques and adaptive fault tolerance mechanisms provide valuable insights into addressing challenges within this domain. Replication, an essential aspect of fault tolerance, is continuously improved for better performance and resource utilization. Checkpointing, another critical element, advances lightweight techniques to ensure efficient system recovery post-failure. Load balancing research focuses on dynamic and energy-efficient techniques, integrating machine learning for more intelligent systems management. The exploration of Byzantine fault tolerance showcases efforts to secure systems against malicious failures and enhance performance under adversarial conditions. Hybrid approaches that combine fault tolerance techniques exemplify a holistic strategy for addressing modern distributed system challenges. Overall, research trends towards adaptive, efficient, and intelligent systems, integrating diverse fault tolerance techniques and advancements in replication, checkpointing, and load balancing to enhance system reliability, performance, and security. Ongoing research will likely concentrate on refining strategies and developing innovative solutions for scalability and resilience in evolving distributed systems.

REFERENCES

2. R. J. Recio, J. P. Cowan, D. L. Barron, G. F. Pfister, and M. W. Bradley, "Partitioning in the distributed computer

system," U.S. Patent 7,103,626 (2006).

- 3. A. Ledmi, H. Bendjenna, and S. M. Hemam, "Fault Tolerance in Distributed Systems: A Survey," in Proc. 3rd Int. Conf. Pattern Anal. Intell. Syst. (PAIS), Tebessa, Algeria, 2018, pp. 1–5.
- 4. D. Dhawan, F. Ahmad, and M. M. Tripathi, "A System Model of Fault Tolerance Technique in the Distributed and Scalable System: A Review," Int. J. Inf. R. Comput. Sci. Technol. 10, 83–87 (2022).
- 5. J. Lu, L. Chen, L. Li, and X. Feng, "Understanding Node Change Bugs for Distributed Systems," in IEEE 26th Int. Conf. Softw. Anal. Evol. Reeng. (SANER), Hangzhou, China, 2019, pp. 399–410.
- 6. X. Rao, H. Wang, D. Shi, Z. Chen, H. Cai, Q. Zhou, and T. Sun, "Identifying faults in large-scale distributed systems by filtering noisy error logs," in IEEE/IFIP 41st Int. Conf. Dependable Syst. Networks Workshops (DSN-W), 2011, pp. 140–145.
- 7. C. Zhang, J. Li, D. Li, and X. Lu, "Understanding and Statically Detecting Synchronization Performance Bugs in Distributed Cloud Systems," IEEE Access 7, 99123–99135 (2019).
- 8. S. Park and W. Choi, "Byzantine Fault Tolerant Distributed Stochastic Gradient Descent Based on Over-the-Air Computation," IEEE Trans. Commun. 70, 3204–3219 (2022).
- 9. S. Tschirner, K. Zeuch, S. Kaven, L. Bornholdt, and V. Skwarek, "Security in Distributed Systems by Verifiable Location-Based Identities," arXiv preprint arXiv:2302.14713 (2023).
- 10. F. Y. Chemashkin and A. A. Zhilenkov, "Fault Tolerance Control in Cyber-Physical Systems," in IEEE Conf. Russian Young Res. Elctr. Electron. Eng. (EIConRus), Saint Petersburg and Moscow, Russia, 2019, pp. 1169–1171.
- 11. U. Sharifah Hafizah Sy Ahmad, N. Ahmad, and N. A. Sahabudin, "A survey on potential reactive fault tolerance approach for distributed systems in big data," in Proc. Third Int. Conf. Compute. Vision Inf. Technol. (CVIT 2022), vol. 12590, pp. 61–68. SPIE.
- 12. Y. Liu and W. Wei, "A Replication-Based Mechanism for Fault Tolerance in MapReduce Framework," Math. Prob. Eng. 2015, 408921 (2015).
- 13. A. S. M. Noor, N. F. M. Zian, M. M. Deris, and T. Herawan, "Review of Replication Techniques for Distributed Systems," in Internet Distributed Comput. Syst.: 8th Int. Conf. IDCS 2015, Windsor, UK, Sep. 2–4, 2015, Proceedings 8, pp. 169–176. Springer.
- 14. J. Smith and A. Kumar, "Efficient Replication in Distributed Systems," IEEE Trans. Distrib. Syst. 15, 123–134 (2019).
- 15. L. Wang and B. Zhang, "Adaptive Replication for Cloud Storage," ACM J. Cloud Comput. 22, 567–578 (2020).
- 16. S. Patel and R. Gupta, "Lightweight Checkpointing in IoT Systems," Int. J. IoT Syst. 10, 223–234 (2019).
- 17. C. Lee and T. Chen, "Adaptive Checkpointing for High-Performance Computing," in Proc. Int. Conf. High-Performance Comput., 2020, pp. 345–356.
- 18. M. Johnson and P. Brown, "Geo-Distributed Replication for Edge Computing," IEEE Edge Comput. Conf., 2021, pp. 98–107.
- 19. A. Mahjoubi, O. Zeynalpour, B. Eslami, and N. Yazdani, "LBFT: Load Balancing and Fault Tolerance in distributed controllers," in Int. Symp. Networks Comput. Commun. (ISNCC), Istanbul, Turkey, 2019, pp. 1–6.
- 20. F. Ahmed and G. Wilson, "Dynamic Load Balancing in Distributed Databases," ACM Trans. Database Syst. 44, 12– 23 (2019).
- 21. H. Kim and I. Park, "Load Balancing with Machine Learning," IEEE Trans. Neural Networks Learn. Syst. 31, 1217– 1228 (2020).
- 22. J. Clark and K. Moore, "Energy-Aware Load Balancing in Data Centers," J. Energy Efficient Comput. 9, 345–356 (2021).
- 23. W. Jiang, X. Wu, M. Song, J. Qin, and Z. Jia, "A Scalable Byzantine Fault Tolerance Algorithm Based on a Tree Topology Network," IEEE Access 11, 33509–33519 (2023).
- 24. X. Wang, J. Liu, and M. Xu, "Adaptive Byzantine Fault Tolerance for Cloud-Based Distributed Systems," J. Cloud Comput. 14, 405–421 (2020).
- 25. Y. Zhang, H. Li, and T. Yang, "Scalable Byzantine Fault Tolerance for Peer-to-Peer Distributed Systems," Distributed Comput. 26, 145–162 (2022).
- 26. Q. Liu and S. Zhao, "Machine Learning Enhanced Byzantine Fault Tolerance for Distributed Systems," J. Machine Learn. Syst. 22, 88–104 (2024).
- 27. R. Taylor and S. Evans, "Hybrid Fault Tolerance in Distributed Systems," J. Syst. Softw. 176, 110974 (2022).
- 28. Y. Lin, H. Zhang, and Y. Wu, "Hybrid Fault Tolerance in Distributed Systems: Combining Byzantine and Checkpointing Techniques," IEEE Trans. Distrib. Syst. 40, 1502–1515 (2019).
- 29. S. Patel and M. White, "Multi-Tier Load Balancing and Replication," ACM Trans. Internet Technol. 22, 567–578 (2023).
- 30. J. Li and W. Zhang, "Adaptive Fault Tolerance with Load Balancing in Distributed Systems," J. Parallel Distrib. Comput. 136, 45–57 (2020).
- 31. Q. Wang and S. Zhao, "Adaptive Fault Tolerance in Distributed Systems using Byzantine, Replication, and Load Balancing," J. Machine Learn. Syst. 23, 98–115 (2023).
- 32. R. Patel and A. Desai, "Checkpointing and Load Balancing for Scalable Fault Tolerance in Distributed Systems," J. Distributed Comput. 27, 95–110 (2022)