APPLYING SOFTWARE ENGINEERING FOR AI/AS ON MISSION-CRITICAL APPLICATIONS IN EDGE COMPUTING

WASP Software Engineering Course Module 2024

Author
Tanaz (Nayereh) Rasouli
Umeå University

1 Introduction

My project topic is ”Resource Management for Mission-Critical Applications at the Edge of the Network”, as it is so abroad, we narrowed it down to ”Fault Tolerance Infrastructure for Mission-Critical Mobile Edge Cloud Applications”. Efficient Mobile Edge Computing (MEC) is about more than just speed; it’s about intelligently deploying limited computational resources where and when they are required the most. Disaster recovery is at the heart of the difficulty, whether a single- edge device fails, partial infrastructure, or the network, receiving crucial data. Considering these crucial data as mission-critical applications, it is a matter of a life/death situation, like healthcare, forecasting, and control of transport systems. Disaster recovery and fault tolerance are crucial elements of resource management in MEC. These components prioritize the maintenance and rapid recovery of edge computing networks in the event of hardware or software failures, especially during critical operations like disaster response. My focus is on a disaster situation like fires or earthquakes which cause damage partially to the edge cluster, how the cluster can still be available while some nodes simultaneously fail, and also the mission-critical applications while their latency is considered. As a use case, we choose weather stations that are damaged by fire. Our work is centered on how to recover this system to ensure that these mission-critical operations are carried out with unrivaled performance and QoS.

Our research investigates the use of current technologies, such as Kubernetes, to effectively handle fault tolerance in situations involving the failure of one or several nodes. In our proposed infrastructure, we employ RabbitMQ as a resilient message broker to ensure dependable message transmission, even during network outages or server failures.

2 Course related

2.1 Robert’s lectures

Based on the lecture 02-feldt-2024-WASP-SE in the testing concept, Verification means ‘Checking whether we are building the product right’ and Validation means ‘Checking whether we are building the right product’ Talking about verification and validation, in both software engineering and AI concepts we need to verify and validate the system from small and simple software to multi-components, and perhaps a distributed system that these components can be independent or dependent on each other. In my project, I can consider service mesh handling with Istio as huge distributed web services that are dependent on each other but as my focus is on mission-critical applications these applications are some critical financial processes concerning business continuity. We need to validate our system not only on the software level but also architecture level to guarantee in the case of peak workload during a disaster and also failures of the network or the physical node the edge cluster is still reliable and application QoS (quality of service) will not violate because in this type of application, the violation consequences is much more critical than any other types of software or system. When considering resource management for mission-critical applications within the context of edge clouds, several ethical concerns arise, particularly regarding data privacy, security, and reliability. In the lecture 04-feldt-2024-WASP-SE, the topic of literature review and snow sampling was discussed, which was close to what I did as a first paper in my PhD. Considering my project, I carried out a survey and submitted it to the 'ACM Computing Journal'. The process of writing a review paper is intricately linked to the type of application considerations done by snow sampling because focusing on the specific type of applications in the Edge area is an open area to research due to the gap. The Research that I did distinguished itself from other forms of information acquisition by its methodical pursuit of novel ideas, prioritizing the creation of original discoveries rather than merely repeating established knowledge. The paper that investigated the lecture focusing on AI claimed that "safety" and "functional correctness" had the highest attraction by the researchers in the "Machine Learning" concept. However, my survey depicted that there is a lack of study on "Fault tolerance" and "security" in the edge computing era.

2.2 Guest lectures

2.2.1 Lenberg, SAAB (guest lecture1)

The lecturer mentioned reasons "why is Software Engineering critical at Saab ATM", one of the reasons was exactly related to my project concerning using the term "safety -critical", ATM systems are safety-critical and need to meet regulatory standards. They can be considered mission-critical applications with less than 100 ms latency. Any failure can cause a human life threat and in the concept of business continuity can lead to millions of dollars lost. In my project, we are trying in case of damages find a solution for vulnerable edge nodes to increase the rate of availability.

The lecturer mentioned that they have redundancy for the nodes and recovery of the downtime should be handled under one second. Reducing downtime is one of the optimization metrics in my project focusing on specific types of application regulatory standards. Although recent researchers focus on "Proactive Methods" by using machine learning to predict failures and try to find solutions before they happen, we focus on more real case scenarios based on disaster recovery in which disasters events like flood or fire are inevitable introducing "reactive methods" and recover the edge nodes as fast as possibles

Talking about the redundancy that is provided in Saab, as they are on-premise data centers is quite different from an edge cluster where nodes are heterogeneous with a low rate of availability and computational capacity.

2.2.2 Dhasarathy, Volvo Trucks (guest lecture2)

The main topic was about using AI and SE on Volvo trucks, in the following paragraph, I explained what I had found related to my project area. Based on the aforementioned applications attribute in my project, applications on autonomous cars can be considered mission-critical applications, they are just mobile edge devices that do not have enough computational capacity to process all the tasks required, so, they will connect to the central base stations and also the other trucks as another edge device. They will upload their tasks to the closest edge servers to process the tasks for them and the results will be back to them but this process should be fast and reliable, any delay can cause choosing the wrong action by the autonomous truck can lead to accidents. The lectures mentioned that AI can be fruitful while testing is essential to be done on the trucks, but it can cause failure in the whole system which has critical consequences. Machine learning applications need a huge amount of memory or computational capacity, they should be hosted in the cloud data center instead of edge nodes which is out of the scope of my project.

3 The CAIN conferences 2023

In this section, two papers related to my project area are chosen and investigated here.

3.1 The first paper

The authors in [1] proposed a "Multi-vocal Literature Review" on both academics and practitioners to investigate design patterns related to Artificial Intelligence (AI) based systems. The number of design patterns in their study is 70, of which 34 are AI-based system modern patterns. The rest is the traditional one, adapted to AI. These patterns are grouped by many aspects of AI systems development, named architecture, deployment, implementation, security and safety, process, testing quality assurance, and topology. The number of patterns of each category is listed below:

- Architecture: 25 patterns
- Deployment: 16 patterns
- Implementation: 9 patterns
- Security and Safety: 9 patterns
- Process: 8 patterns
- Testing and Quality Assurance: 5 patterns
- Topology: 3 patterns

They designed a web application to recognize the viewable patterns. The web-based pattern repository can help researchers and practitioners to access it easily. It can also be fruitful for facilitating the finding of relevant design patterns much more efficiently. The authors mentioned that the way a pattern can be described, each of them should have some attributes recorded, such as the pattern name, the reason for saving it, both negative and positive consequences, and the sources that used the patterns. They believed these records' goals would be effectively understanding and applying the patterns.

Considering AI-based systems, these systems can sometimes have issues moving from development to production. As the article provided a set of proven design patterns, this can be used for AI-based organizations to notice common problems and enhance AI-system in terms of robustness, scalability, and maintainability. To continue, it should be considered that the design patterns can enhance the quality of the software by improving the applications' attributes named reliability and throughput. The software quality to provide such attributes is essential for AI-based systems.

Another essential aspect is that the repository and its related documents are considered a knowledge base. This concept can exceed the learning curve process and improve AI system developments.

Considering the question “How the paper relates to your research”, I am not sure if I can point it out correctly or not but my research on fault tolerance and disaster recovery in mission-critical cloud-native applications can leverage the design patterns identified in this paper. We can improve the reliability, and efficiency of the system by adapting these patterns to the unique challenges of MEC. In this case, we can ensure the continuous operation of these applications even in the face of partial infrastructure failures. As authors in [2], mentioned they have a repository in a machine-readable format, and each of them has a selected JSON, which means all patterns are converted to a JSON format as a part of the repository in all separate files. In my point of view, the mentioned paper identified design patterns like architecture, implementation, and quality assurance that can help me to design and implement a proper fault-tolerance mechanism in my existing infrastructure. The infrastructure can guarantee the availability of mission-critical applications in case of disaster recovery where my physical infrastructure is partially damaged some nodes are failed and are not recoverable. Especially considering architectural patterns that guarantee high availability can be essential for my proposed infrastructure. If we designed a layered architecture for the MEC environment, the individual node failures cannot have an impact on the overall system per

I want to combine question C into on parts. First, ”how the paper could help improve the project”, then I will answer the next part in the heart of the explanation for this question. In my point of view, the idea presented in the paper ”Workflow Pipeline Pattern” can be beneficial. This pattern uses tools like TFX or Airflow to create a pipeline that gets the data, trains the model, and deploys it in one step. This makes the workflow more reproducible and documented. The Feature Store pattern decouples the feature engineering step from model training. As I explain in its definition, it can be helpful to have a scalable pipeline to manage components related to machine learning. It provides a containerized environment that can capsulate each phase of machine learning named data preprocessing, training, and prediction in a container as one of the advantages of using containers is every dependency or config does not have any side effects in the others. By taking benefit of REST API containers are orchestrated and connected to each other while they are independent. Every interaction between containers is efficiently provided by REST APIs. In disaster situations, applying this pattern, I believe can be fruitful because it can improve the system’s performance, reliability, and scalability. The reasons behind this can be considered as follows:

- due to containerizing each step of the pipeline, as I explained before, we have an isolated environment that allows every update or maintenance of individual components not to have any effect on the performance of the whole system or the other components. In this case, replacing or upgrading can be done easily. For example, if I develop a new prediction for the project in a disaster situation without any downtime which is an important factor in the disaster recovery techniques, it can be integrated into the pipeline.

- Considering the scalability, one of the most important features of containerized orchestration for example Kubernetes which I am using on my project is that every component can be scaled in or scaled out based on the demands. The system can scale up the prediction model containers if loads of data increase to ensure the accuracy of the prediction. Not only it can have impact on the performance of each module of machine learning as a container, but also this trend is necessary to have an efficient resource allocation in disaster situations when many systems are damaged or we have power outages. Kubernetes can be seamlessly added into process pipelines and AI pipelines, guaranteeing efficient task rerouting and workload management, even in the event of failures. In this way, I can develop AI pipeline to develop a strategy to keep the whole edge cluster functionality while some edge nodes are failed due to the damage of catastrophic events.

- In my project I am using RabbitMQ to ensure the reliability of the data that the system receives from IoT devices, it can be said, that the nature of the pipeline will help to enhance fault tolerance, it can be restarted or replaced easily while ensuring the system continuous operations, so it can enhance the reliability which is one the major metrics for disaster recovery and fault-tolerance techniques. With RabbitMQ, I can take the benefit of a rule-based safeguards pattern to guarantee the integrity of messages as they are processed and consumed through disasters

I have weather stations as my use case in my project (mission-critical systems and applications), I can apply the AI pipeline pattern to ensure that data from a partially damaged edge cluster and a weather station is processed while breaking each into smaller and manageable component with critical messages as data from IoT devices and cameras in the disaster.

In the end, the research has the potential to significantly enhance the artificial intelligence engineering aspects of the project if it incorporates and puts into practice the design patterns that are described in the paper. The modularity, scalability, stability, and ease of maintenance of the system will all be improved as a result of these adjustments. We can take advantage of using this strategy not only by simplifying the procedures for development and deployment but also by improving the efficiency and durability of mission-critical applications in emergency scenarios. This would ensure that the system can quickly adapt to new challenges and demands.

3.2 The second paper

The authors in [2] benefit from existing machine learning techniques, “Federated Learning” to enhance the mutual benefit of data-sharing ecosystems. They help those organizations that can share machine learning tasks with other organizations without sharing critical data directly. They preserve data privacy, and security, and enhance the data-sharing ecosystems. Each organization trains its models on its local data and just shares the parameters of the trained model. Their proposed method was applied to optimize a city traffic prediction algorithm, so, their method will work for multi-systems with data-sensitive sharing.

Based on the definition of regulation in [3] it refers to the deliberate and concentrated effort to change the actions of individuals based on specific criteria or objectives, aiming to achieve a well-defined result or results. This process may involve establishing rules, collecting information, and modifying behavior. Therefore, we can consider the ethical aspects of using data in terms of privacy and security as one the important aspects of AI after a wide range of usage of machine learning algorithms, which is considered by the paper. Data privacy is provided by Federated Machine Learning which allows the use of shared data without exposing private information related to each organization. On the other hand, it can help organizations leverage more extensive datasets collectively which can improve the machine learning models.

Regarding my project, as I am working on a fault-tolerance infrastructure for mission-critical applications in the edge of the network, I believe this paper can be related to my project in ensuring data availability and integrity of data during failures in the edge cluster via the disaster events. They applied their model to optimize the traffic prediction model, as the control of transport is one of the mission-critical applications and use case we can say the traffic model kind of related to control transport systems. Another important aspect is that my project involves real-time data management -critical data- the same as the traffic prediction model which should also deal with real-time data management.

Integration into a larger AI-intensive software project focused on disaster recovery and fault tolerance for mission-critical applications, I think we can apply the idea of the paper to enhance the project's capabilities. The improvement could be as follows:

- **Enhancing the privacy and security**
  In disaster management systems -in my project's use case different weather stations- multiple systems should share critical data where the paper idea can be applied to guarantee the preservation of critical data. Data privacy in mission-critical applications such as healthcare and emergency response scenarios is a significant factor, so, applying the idea of the paper can minimize the risk of exposing sensitive information.

- **Improving data utilization**
  Based on the previous item explanations, as federated learning can enable the integration of diverse datasets from various clusters, this secure variety of datasets can lead to better protection models in case of catastrophe in disaster recovery.

In my point of view, if we apply federated learning to edge computing, it can lead to training a much better model than without edge, the reason is that the federated learning algorithm can collect more datasets from the different edge nodes on the distributed clusters, so, it can make a very accurate prediction model. An accurate model is essential during disaster recovery operations.

References

[1] L. Heiland, M. Hauser, and J. Bogner, "Design patterns for ai-based systems: A multivocal literature review and pattern repository," in *2023 IEEE/ACM 2nd International Conference on AI Engineering–Software Engineering for AI (CAIN)*. IEEE, 2023, pp. 184–196.

[2] I. Krasteva, B. Kraychev, and E. Kiyamousavi, "How federated machine learning helps increase the mutual benefit of data-sharing ecosystems," in *2023 IEEE/ACM 2nd International Conference on AI Engineering – Software Engineering for AI (CAIN)*, 2023, pp. 96–97.

[3] M. Hildebrandt, "Algorithmic regulation and the rule of law," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2128, p. 20170355, 2018.