Influence of Artificial Intelligence on Microservices Architecture

Navjot Saroa UCID: 30193854

I. Introduction

The purpose of this research is to discuss how Artificial Intelligence (AI) can potentially influence the microservices architecture (MSA). MSA is an improvement on the monolithic architecture, where applications are broken down into smaller components that perform specialised tasks, as opposed to the traditional approach of the monolithic architecture, where all the functionalities are in a singular application [1].

This focus on splitting applications into smaller parts makes MSAs more modular. Since each microservice operates independently; their development, deployment, and scaling does not affect other microservices. [2]. This has resulted in the architecture becoming increasingly popular over time.

However, these features bring in their own challenges. This paper will focus on its issues by describing these problems, focusing on resource allocation and security concerns. Following that, the paper will discuss how AI can be used to address these issues. Finally, the paper will conclude with the problems the use of AI itself brings, and whether its use is therefore worth the costs it comes with.

This research is especially important due to the popularity of MSAs as the architecture of choice as 96% of all applications today are built on it [3]. Furthermore, the nature of the problems MSAs face means that AI lends itself particularly well to architectures like these [4]. As the demands on these architectures keep increasing; static, rule-based solutions will no longer be sufficient [5]. As a result research into AI based dynamic solutions will become more crucial with time.

II. Goal and Research Question

This study will highlight the proposed use cases for AI to improve the microservices architecture and discuss the potential benefits and pitfalls of using AI in this context.
The specific question being answered is ”how AI can be used to improve microservices?”.

III. Method

This paper is a rapid review, which is the systematic analysis of academic research, summarising key findings briefly. Rapid reviews are done with the aim of providing a summary of the research done on a particular topic, which is the influence of AI on microservices in this case. In this process, parts of the systematic review process are not included so that information is presented quickly and in a concise manner. [6]

For this paper, a rapid review was carried out on 15 papers, these papers were found on Google Scholar. Search terms such as ”AI”, ”microservices”, ”optimisation”, ”security”, and ”sustainability” were used to find the relevant research. The search was filtered to only present papers that were published within the past 10 years (2015 to 2025).

Snowballing was also used to find papers. This is the process of discovering research through the references of papers that have already been found [7]. The same idea of filtering out papers published before 2015 was used. Paywalled papers were also excluded from this search, except for the ones that could be accessed through the institution of University of Calgary.

The primary concepts searched for within each paper were concrete ideas of how AI could be used in this given context, with technical elaboration on how those ideas could be implemented. The reader will notice that Ramamoorthi’s papers were especially useful in this regard.

For the final section of the paper, where the challenges of using AI are discussed, the search was more general. The search started off by looking purely at the environmental impact of AI. However, as time progressed, other challenges that are relevant to this paper’s content were found, so they were included as well.

IV. Results

The two primary challenges that were brought up in the papers reviewed were of the lack of efficiency in resource allocation, and the increased security risks when MSAs are based on static rules and algorithms:

A. *Efficiency*

Since this architecture results in complicated interdependences between microservices, there is an issue of resource contention and bottlenecks in performance [8]. This complexity also results in a higher consumption of network and computing resources [9]. Therefore, Richardson et al. suggest that optimisation algorithms can be used to mitigate the effects of these problems [4], using Reinforcement Learning (RL) to optimise the placement of microservices.

An agent responsible for the allocation of network and computational resources can learn the best course of action to take by getting feedback (rewards or penalties) on its interactions with the environment it is in [10]. Doing so not only optimises the allocation, but also makes the architecture inherently dynamic, since resource distribution is no longer determined through static rules, and is instead constantly changing based on the feedback it receives. [4]. This results in lowered costs, lower power consumption, and lower latency [3].

Additionally, Ramamoorthi suggests using Predictive Analytics (PA) and Evolutionary Algorithms (EA) for this. PA bases its decisions on historical data like patterns in resource usage (certain microservices use more resources), time of day (demand is higher during the day), etc [11]. EA uses a more exploratory approach, evaluating a large group of possible configurations, choosing the best performing ones for reproduction. After sufficient generations, an optimal configuration is found and can be implemented [12], [13].

Richardson et al. also discuss the concept of multi-objective optimisation in particular amongst even more techniques [4], these kinds of problems are well suited to AI as well. Ramamoorthi fleshes out this idea, identifying the cost of deploying microservices and the latency of the system as the two factors that must be minimised. Specifically, a linear combination of functions of cost and latency of each microservice on each edge node should be minimised, whilst keeping latency below a specific limit [5].

Figure 1 shows how AI compares to the static, rule-based scaling by Kubernetes Horizontal Pod Autoscaler (HPA) in various types of traffic. AI consistently outperforms Kubernetes HPA, having lower latency in all situations, and a higher throughput.

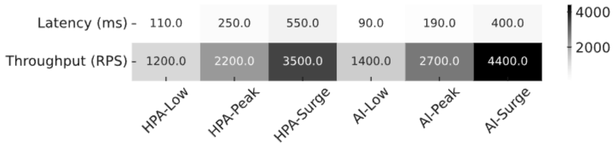

Fig. 1. Latency (ms) and throughput (requests per second) compared between Kubernetes HPA and AI, in low traffic, high traffic , and unpredictable traffic surge situations [12]

B. Security

Since, by nature, applications are very loosely coupled and highly modular [2], yet very interconnected, a security failure in just one microservice could result in a compromise of the entire application [14]. Although Al-Doghman et al. specifically talk about microservices in the context of Internet of Things (IoT), the idea that malicious users can interact with individual nodes (or in the context of this paper, microservices) still stands. Therefore, it is important to monitor these interactions and distinguish between normal and abnormal interactions. Al-Doghman et al. suggest that Machine Learning (ML) and Deep Learning (DL) can be used for situations like these due to their classification abilities. Once a model has been trained on what normal interactions look like, it can detect and flag any abnormal interactions. They also proceed to suggest that DL can be used to predict future attacks that are variations of previous ones [15].

Ramamoorthi proposes the use of AI for intrusion detection, combined with an automated mitigation engine, to identify and conteract malicious actors that could take advantage of the distributed nature of microservices. They suggest that AI can be used for dynamic firewall management (dynamic adjustment of firewall settings) and service isolation and recovery (where the RL agent can isolate compromised microservices and restore its integrity after the intrusion has been handled) [16].

Figure 2 depicts the usage of resources across different phases of an intrusion, illustrating the benefits of service isolation and recovery.

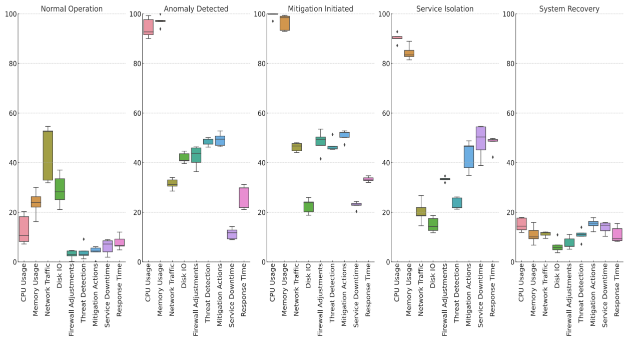

Fig. 2. Resourse usage metrics through the process of service isolation and recovery [16]

C. Challenges of Using AI

Despite the potential benefits that come with the use of AI, it is important to consider whether or not the costs of using AI are indeed outweighed by its benefits. Figure 3 shows the carbon footprint of AI compared to other activities, it is clear that the carbon footprint of training an AI model is several times higher than what the average human would produce in a year (113.85 times greater). Considering the populariity of microservices (as mentioned before, 96% of the applications today use it [3]), implementing AI technology to mitigate its challenges could in turn result in larger challenges from a sustainability point of view. Data centres, facilities that help in the running of major AI models, are projected to have a 15-fold increase in energy demand by 2030, amounting to 8% of the projected global demand [17].

Other disadvantages of using AI include interpretability, and difficulty in training [19]. The predictions made by an AI can sometimes be difficult to understand, and therefore, hard to justify, to trust, and to maintain. Furthermore, a significant amount of data is required to effectively train an AI model in

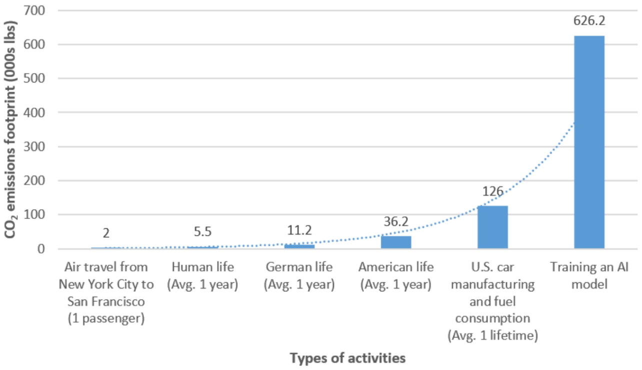

Fig. 3. Carbon footprint of AI compared to other types of activities. ”Carbon Footprint of AI = Carbon Footprint of Computing Technology Production + Carbon Footprint of Training AI + Carbon Footprint of Operating AI (Usage)” [18]

the first place, this may not be something all developers may have access to.

These challenges must be addressed before AI can be considered to be a long term solution to the issues MSAs face.

TABLE I
SUMMARY OF RESOURCES USED

Reference	Usage
[3]–[5], [8]–[10], [12]	Discussion regarding AI’s impact on the efficiency of MSAs.
[2], [14]–[16]	How AI can help make MSAs more secure
[17]–[19]	Downsides of using AI
[11], [13]	Exclusively used for definitions of terminology in the results section

V. CONCLUSION

The study identified that the key challenges for microservices are those of efficiency and of security. AI is capable of solving these issues through the use of optimising algorithms, and using ML and DL for identificaiton of security threats. However, the magnitude of the environmental impact of training and using AI is such that it must be taken into account when deciding to use it, because the costs of using AI may outweigh the benefits it brings.

Two areas of research are recommended as a result of this study. First, development of sustainable AI models is cruicial to ensure that AI can be a long term solution to the issues all software architectures face. Second, research into the influence of AI into other software architectures could prove to be useful. For example, the microkernel architecture behaves in a similar way to microservices, it facilitates the loading of the required plug-ins as required in order to improve power consumption and enable dynamic activation of features. [20]

Generative AI disclosure: ChatGPT and Github Copilot were used in the building of the structure of this LaTeX document, and to improve the wording of the ideas mentioned in this paper. No generative AI was used as a research tool to gather resources or generate ideas for the paper.

REFERENCES

[1] L. De Lauretis, “From monolithic architecture to microservices architecture,” in *2019 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, 2019, pp. 93–96.

[2] J. I. Akerele, A. Uzoka, P. U. Ojukwu, and O. J. Olamijuwon, “Improving healthcare application scalability through microservices architecture in the cloud,” *International Journal of Scientific Research Updates*, vol. 8, no. 02, pp. 100–109, 2024.

[3] P. Keshavarz Haddadha, M. H. Rezvani, M. MollaMotalebi, and A. Shankar, “Machine learning methods for service placement: a systematic review,” *Artificial Intelligence Review*, vol. 57, no. 3, p. 61, 2024.

[4] N. Richardson, S. Kothapalli, A. R. Onteddu, R. R. Kundavaram, and R. R. Talla, “Ai-driven optimization techniques for evolving software architecture in complex systems,” *ABC Journal of Advanced Research*, vol. 12, no. 2, pp. 71–84, 2023.

[5] V. Ramamoorthi *et al.*, “Real-time adaptive orchestration of ai microservices in dynamic edge computing,” *Journal of Advanced Computing Systems*, vol. 3, no. 3, pp. 1–9, 2023.

[6] B. Cartaxo, G. Pinto, and S. Soares, “Rapid reviews in software engineering,” *Contemporary Empirical Methods in Software Engineering*, pp. 357–384, 2020.

[7] S. Jalali and C. Wohlin, “Systematic literature studies: database searches vs. backward snowballing,” in *Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement*, 2012, pp. 29–38.

[8] B. Barua and M. S. Kaiser, “Ai-driven resource allocation framework for microservices in hybrid cloud platforms,” 2024. [Online]. Available: https://arxiv.org/abs/2412.02610

[9] J. Soldani, D. A. Tamburri, and W.-J. Van Den Heuvel, “The pains and gains of microservices: A systematic grey literature review,” *Journal of Systems and Software*, vol. 146, pp. 215–232, 2018.

[10] D. Ernst and A. Louette, “Introduction to reinforcement learning,” *Feuerriegel, S., Hartmann, J., Janiesch, C., and Zschech, P*, pp. 111–126, 2024.

[11] V. Kumar and M. Garg, “Predictive analytics: a review of trends and techniques,” *International Journal of Computer Applications*, vol. 182, no. 1, pp. 31–37, 2018.

[12] V. Ramamoorthi *et al.*, “Ai-enhanced performance optimization for microservice-based systems,” *Journal of Advanced Computing Systems*, vol. 4, no. 9, pp. 1–7, 2024.

[13] D. W. Corne and M. A. Lones, “Evolutionary algorithms,” *arXiv preprint arXiv:1805.11014*, 2018.

[14] Y. Sun, S. Nanda, and T. Jaeger, “Security-as-a-service for microservices-based cloud applications,” in *2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom)*. IEEE, 2015, pp. 50–57.

[15] F. Al-Doghman, N. Moustafa, I. Khalil, N. Sohrabi, Z. Tari, and A. Y. Zomaya, “Ai-enabled secure microservices in edge computing: Opportunities and challenges,” *IEEE Transactions on Services Computing*, vol. 16, no. 2, pp. 1485–1504, 2022.

[16] V. Ramamoorthi, “Anomaly detection and automated mitigation for microservices security with ai,” *Applied Research in Artificial Intelligence and Cloud Computing*, vol. 7, no. 6, pp. 211–222, 2024.

[17] N. Jones *et al.*, “How to stop data centres from gobbling up the world’s electricity,” *nature*, vol. 561, no. 7722, pp. 163–166, 2018.

[18] N. E. Mitu and G. T. Mitu, “The hidden cost of ai: Carbon footprint and mitigation strategies,” *Revista de Științe Politice. Revue des Sciences Politiques• No*, vol. 84, pp. 9–16, 2024.

[19] A. Singh and A. Aggarwal, “Artificial intelligence enabled microservice container orchestration to increase efficiency and scalability for high volume transaction system in cloud environment,” *Journal of Artificial Intelligence Research and Applications*, vol. 3, no. 2, pp. 24–52, 2023.

[20] K. Nandy, S. SM, A. Bhadauria, and S. Upadhyay, “Resource optimization in edge through microkernel architecture,” in *2024 IEEE 21st International Conference on Software Architecture Companion (ICSA-C)*, 2024, pp. 362–367.