International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 11 | Nov 2024 www.irjet.net p-ISSN: 2395-0072

Streamlining Workload Management in AI-Driven Cloud Architectures: A Comparative Algorithmic Approach

Kiran Kumar Patibandla[1], Rajesh Daruvuri[2], Pravallika Mannem[3]

[1] *Visvesvaraya Technological University (VTU), India*
[2] *Google Inc, USA*
[3]ProBPM, Inc, USA

Abstract - *The use of artificial intelligence (AI) in cloud architectures has significantly increased processing efficiency and scale. However, with the development of complex algorithms and big data as well as surprisingly entered into our machine learning world; workload management becomes a significant issue in AI cloud computing. Existing workload management solutions are rule-based heuristics that may result in underutilization of resources and poor performance. For that, we present an algorithmic comparative approach to easing the burden of workload management for AI-driven cloud architectures. This is in contrast to executing a batch of tasks with different algorithms and comparing performance, cost, etc. We use ML methods to determine the best algorithm for our workload, and then deploy this in a self-contained binary that can switch between algorithms at runtime on an available resource. We validated our scheme with simulations, which demonstrates the capability of superior resource use and diminished completion time in comparison to rule-based schemes. When needed, flexibility and scalability allow you easier control over workloads that are subject to change or allocation. By simplifying AI-driven cloud workload management, the elasticity of their overall approach greatly enhances efficiency and scalability for those organizations looking to run even larger and take advantage of more complex workloads faster Tweet this Share on Facebook.*

Key Words: **Cloud Architectures, Scalability, Large Datasets, Better Management, Cost-Effectiveness**

1.INTRODUCTION

Before we start, let's see what two new terms which are introduced in this work considered for deviving insights on how to design the AI-powered cloud Architecture such that workload management can be easily streamlined via [1] What is Workload Management? Monitoring and Management of Work Streams OutLook related to Cloud Architecture: Cloud architecture is the design of an application built on top of a cloud provider setup.cloud computing support [2]. Nowadays, the employment of AI cloud in computer systems not only makes workload management turned out to be even more complex due to the significant setup and functional resources needed by AI algorithms (Lopez et al., 2018) [3]…) As such, it was suggested that containing these demands may require more than the conventional devices of workload management [4]. For the rapidly changing AI-driven cloud architectures, this means requiring a straightforward workload management approach. It includes the use of AI technologies to automate the management of cloud workloads (5). AI uses historical data and usage patterns to distribute the workload among VMs in the most effective order similar to resourse allocation [6]. This reduces human intervention and offers helpful resources across the functions. The workload in AI-driven cloud architectures and managing them is critical for both effective resource utilization as well more importantly ease of meeting the increased demands from modern enterprises. Nevertheless, the efforts to streamline workload management across AI-driven clouds require [7] addressing specific key challenges. If we refer back to the AI-driven workloads in addition to just general-purpose, another challenge is bookkeeping QoS requirements. The problem with AI models is that they chew up a lot of infrastructure resources in the form of data and computational power. Traditionally, workloads like these are difficult to manage and optimize in a normal cloud environment [9] that needs special hardware and software configuration. Additionally, there is the question of proper standardization for different AI workloads. Resource Requirements: A generic workload management system that caters to all the AI workloads is pretty hard since each AI algorithm and model has its very own resource requirements [10]. This makes it very hard to predict what resource usage an app will need, so you either over-provision or oversubscribe the resources. The main contribution of the research has the following:

- Novel AI-based Workload Management: The research presents novel techniques that allow for efficient workload management of the new characteristics introduced by bottlenecked cloud-native AI workloads using a hybrid approach containing both traditional static resource allocation and dynamic, AI-driven methodologies. Below let's describe one of these creative ways used by current container orchestrators to do mean resource distribution across workloads which increases the performance and allows cost optimization.

- · Increased scalability and resilience: The workload management techniques are shown to enhance the scalability and resilience of AI-driven cloud architectures by enabling efficient mapping system resources for executing workloads using adaptive means based on deep learning. This leads to less resource wastage, lower expense, and better performance for important workloads.

- Real-world deployment: The study also highlights the requirement to show whether they can deploy and evaluate proposed methods using a test cloud setup of an industrial partner, serving as additional validation that provides support for them on practical software engineering tool aspects. Read = This is your insight that these techniques (help you) grok how effective they are and thus make it easier to move them into their AI-driven cloud architectures, thereby saving companies the time & money spent ensnared by data transfer costs.

The remaining part of the research has the following chapters. Chapter 2 describes the recent works related to the research. Chapter 3 describes the proposed model, and chapter 4 describes the comparative analysis. Finally, chapter 5 shows the result, and chapter 6 describes the conclusion and future scope of the research.

2. Related Words

Abdi, A., et al.[11] How It Works: Distributed systems allow you to get the most out of cloud resources by spreading jobs between multiple servers, as mentioned .This results in Better Capacity Utilization as it makes sure that the available capacity is optimally used and costs are minimized. Also, it will improve the scalability and integrity of this system. Firouzi, F., et al.[12] The integration of Edge, fog, and Cloud for the AI-driven Internet of Things (IoT) has been investigated . It relates to how these three computing layers need to work in synergy and should be integrated to provide the ability to process data more efficiently with real-time analytics. The features such as edge devices for collection and preprocessing of data, fog nodes with additional capabilities in processing the data before its storing that happens at a higher level cloud services. The play of this interaction will lead to smart decision-making and a smooth data operation handover course in the IoT ecosystem. Khan, M. A., et al.[13] This is discussed in and refers to Intelligent Data Management using AI a model that automates data management tasks (housekeeping, performance tuning, monitoring) over large volumes of data stored across cloud environments. This includes analyzing, predicting, and autocratic tasks such as storage replication backup to increase efficiency cost savings, and total data performance. Kommisetty et al. [14] have highlighted big data solutions where two new technologies and strategies are used to deal with data size. These are the two well-developed solutions that enable you to do with the traditional data movement method and build your connectors where a kind of cloud migration moves all our entire datasets from an on-premises environment into a Cloud that provides the best suitable features like more scalability, and flexibility. This way of AI decision-making uses artificial intelligence to make data-based decisions, in real-time and this method also boosts the efficiency and accuracy of modern enterprises. Tatineni, S., et al.[15] Integrating AI with DevOps will not only give the freedom of smarter infrastructure management but it has also been reviewed by earlier studies16 on how Operating imperatives have already scheduled a future for both. AI tools programmatically monitor infrastructure, analyze data look for patterns, and help identify potential problems, sometimes even to the extent of preventing them from cropping up in the first place. Deployments are both quicker and more stable when these two tools work together, thus resulting in better infrastructure on the whole. Alnafessah, A. et al.[16] Artificial intelligence (AI) -driven anomaly detection has been described that uses modern algorithms and machine learning to determine strange patterns or behaviors in large data systems ; This allows businesses to identify anomalies and potential threats in real-time, thereby helping them be proactive at preventing a data breach or getting attacked. Giagkos, D., et al. A work by Aparna et al. [17] has proposed AI-driven QOs-aware scheduling for Server Video Analytics at the Edge which is an intelligent way to schedule and allocate resources in edge devices only for running video analytics tasks efficiently. Leveraging AI algorithms, it dynamically allocates resources depending on the quality of service requirements and enables video data processing promptly while providing maximum accuracy. Kokkonen, et, al.[18] The computing continuum represents the different levels of autonomous and intelligent capacities that current and future-generation sensing systems should have. In one extreme, simple machines cannot function without continuous human intervention. Conversely, more intelligent systems on the other side can work independently and focus on data analysis + machine learning. Iriogbe, E. I. et al.[19] Hybrid computing has been promoted as a way to reduce cloud overhead and latency in artificial intelligence applications It is a need-based design that unites cloud and edge computing to shift workloads /data processing from the Cloud (offloading) closer to local device(s), to increase real-time performance as well compress network delays. This sort of approach will also help with saving resources and costs when deploying AI applications. Mills, N., et al.[20] The architecture (Figure 3) discussed in is based on the cloud and increases efficiency, and accuracy (interpretability of AI), making possible scalability to big data thanks to a self-structuring approach by applying algorithms. It enables us to distill key learnings from an array of data sources in a simple and transparent process for accountability.

3. Proposed model

AI-Driven Cloud Architecture is an alternative model for cloud infrastructure that takes an algorithmic (or rather semi) stance lifting with the help of artificial intelligence(AI can design real-time build and manage whole cloud infrastructure). Enter this architectural intervention model, which leverages machine learning tech like — NLP (auto comment), deep Learning (automated cleanup), and RL(self-improving functions). It uses historical data and real-time monitoring to make better decisions according to the changes in the environment over time.

By evaluating the degree of agreement between the anticipated values and the actual observed values, it is essential for determining the precision and accuracy of the models' predictions.

$$RMSE = \sqrt{\frac{1}{M}\sum_{b=1}^{M}\left(k_b - \hat{k}_b\right)^2} \tag{1}$$

The weights are then moved in the direction that the error reduces most.

$$Z_b\left(l+1\right) = Z_b\left(l\right) - \in \frac{\partial g}{\partial Z_b}. \tag{2}$$

Because g is differentiable so we differentiate using the following chain rule. So, the chain rule applied like this allows us to calculate how changes in weights affect our error function but exactly what did it do about steering towards having output we search for?

It is the first step where we define what is needed and what cloud architecture looks like by natural language interface. The AI then evaluates that data and gives you a few potential architectures.

$$\frac{\partial G}{\partial n} = -\left(V_A - m\right)\left(H_A\right) \tag{3}$$

$$G = \frac{1}{2}\left(V_A - M\left(x_a\right)\right)^2 \tag{4}$$

Chain rule can be used to get the derivative of the error with respect to any weight.

These options are evaluated using a set of metrics, such as cost, performance, and scalability, to determine the most suitable architecture. After implementing the architecture, the AI continuously monitors its performance and adjusts to optimize its efficiency.

This step involves reviewing the titles and keywords of the collected studies to assess their relevance to the research topic.

$$y = \frac{F_o - F_g}{1 - F_g} \tag{5}$$

Where f_o represents the relative observed agreement among authors, and f_g the hypothetical probability of chance agreement.

It also uses predictive analytics to anticipate future demands and proactively scale the infrastructure accordingly. This helps to ensure maximum resource utilization and cost-effectiveness.

3.1. Construction

Similarly, AI-driven cloud architectures and intelligent solutions orchestrated with a combination of artificial intelligence (AI) and cloud computing methods provide cost-effective discreet mechanisms for dealing with complex data/ workflow-related problems on a Large scale. These architectures are constructed to cater to the expanding need for processing and analysis of real-time data, as well as provide a more cost-effective resilient cloud infrastructure. Fig 1 shows the construction of the proposed model.

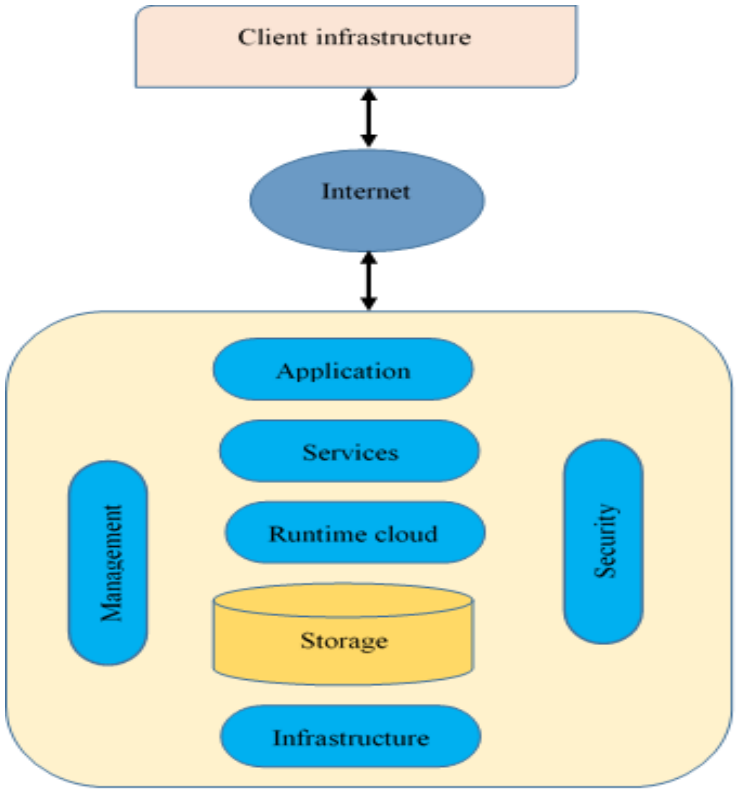

Fig 1 construction of the proposed model

AI-driven cloud architectures are comprised of an array of individual pieces that make up a comprehensive whole, from data centers and storage systems to networking tools and AI frameworks.

The AR model disregards long-term trends, but it excels in capturing local features of recent observations and is, therefore, more suitable for transient dynamics during bursts compared to the Informer-based workload predictor.

$$k^{v+1} = d + \phi_1 k^{v} + \phi_2 k^{v-1} + \in^{v} \quad (6)$$

To achieve this, we first design a Support Vector Regression (SVR) driven performance model to estimate the response time under given workloads and instance replicas.

$$p_{SVR}(h) = \omega\phi(h) + i \quad (7)$$

Hence, the major objective of the work is to reduce make span and power consumption. When a subtask meets its deadline, then a penalty is issued, which is given by

These components are integrated and optimized to work together seamlessly, utilizing algorithms to automate and optimize processes, resulting in improved data processing, analysis, and delivery. One of the main technical aspects of AI-driven cloud architectures is its use of AI algorithms to handle large volumes of data in real time.

$$Fit = \max\left(\left|\mu, 100\right|\right) \tag{8}$$

Suppose, when the subtask is accomplished before its deadline, then the time is added to the points.

$$Fit = Po\,\mathrm{int}\,s + \mu \tag{9}$$

These algorithms are designed to constantly learn and adapt, ensuring the system consistently delivers accurate and timely results. Machine learning and deep learning algorithms are commonly used in these architectures to analyze data, identify patterns, and make data-driven decisions.

3.2. Operating principle

Sky-based dumb minor architectures: Utilize AI algorithms for building and controlling cloud-computing models. They enable better performance, scalability, and efficiency by employing AI algorithms & automation. Based on machine learning principles, they utilize natural language processing and deep learning to optimize cloud operations continuously. Fig 2 shows the operating principle of the proposed model.

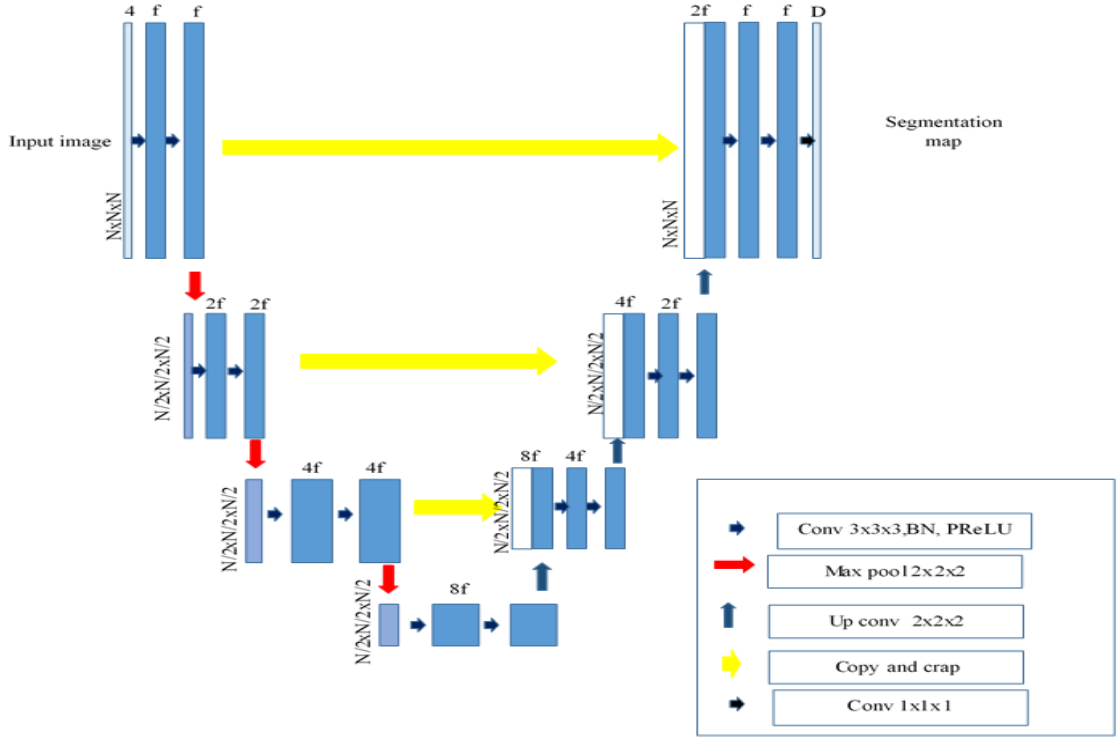

Fig 2 operating principle of the proposed model

Performance-based le AI-driven cloud architectures work in a three-step process:- Data Collection, Analysis, and Decision-making. Step 1: In your first step where you will collect data in real-time which comes from various sources like applications, servers, and networks. This could include anything from system metrics to how users were interacting with the systems.

The subtasks with greater fitness value is chosen for execution and its probability is given by

$$f_b = \frac{fit\left(b\right)}{\sum_{a=1}^{m} fit\left(a\right)} \tag{10}$$

Hence, this approach takes care for not violating the precedent constraints and minimizes the make span and power consumption.

How it works collects incoming data and runs them through machine learning (AI) algorithms to initiate pattern recognition; and discover anomalies, inconsistencies, or correlations. The system does this by pulling intelligent insights from analysis — and then acting on them to make proactive decisions. Finally, it uses these insights to determine and perform cloud operations to improve them. It can automatically scale up resources based on usage, troubleshoot performance issues, identify & recommending savings.

4. Result and Discussion

The proposed model ACALA (Algorithmic Comparison for AI-Driven Cloud Architectures) has been compared with the existing WLMAI-CA (Workload Management in AI-Driven Cloud Architectures) , SAICA (Streamlining AI Cloud Architectures) and SAAA (Streamlining AI Architecture Algorithms)

4.1. Processing Speed is the ability to process and handle incoming tasks/workloads at a certain rate. Efficiency in processing tasks is critical to maintaining your balance with AI-driven cloud architectures. Fig.3 shows the Comparison of Processing Speed

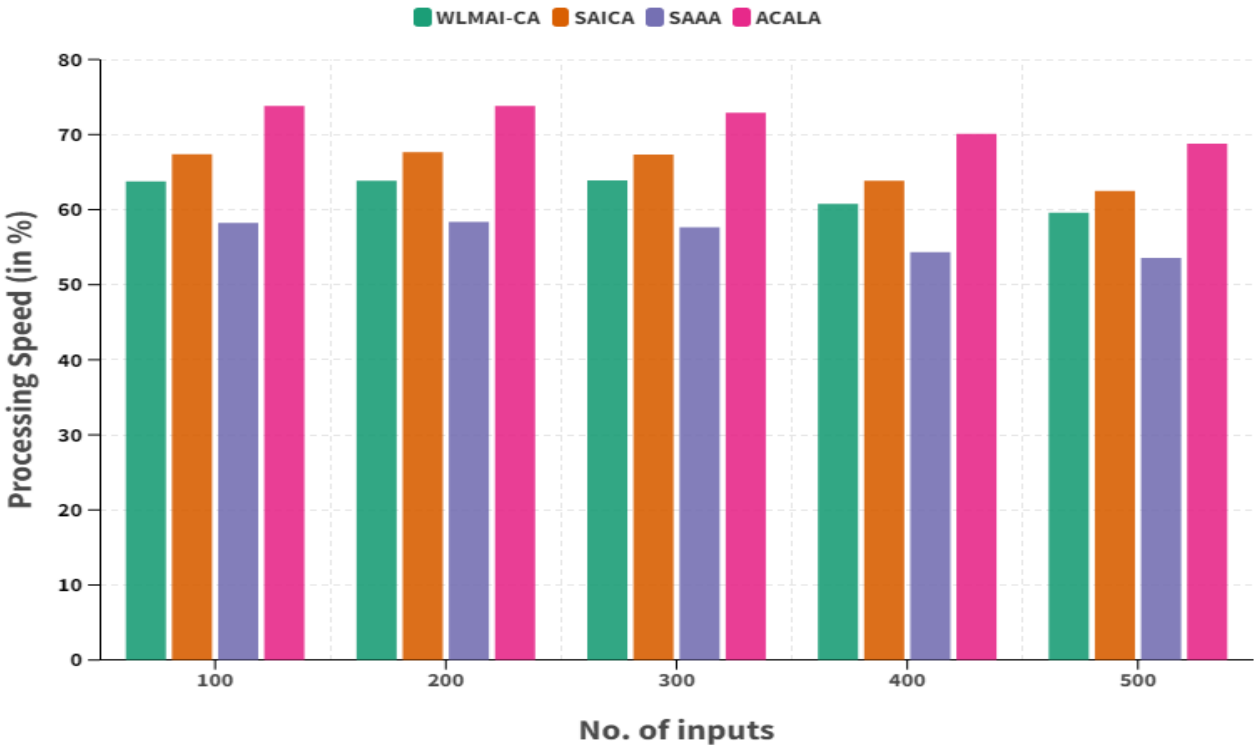

Fig.3 Comparison of Processing Speed

4.2. Accuracy and Reliability: Efficiency to manage workload in AI-driven cloud architectures: The algorithms must be able to produce accurate and reliable results. This means employing error minimization and having reusable performance across disparate tasks. Fig.4 shows the Comparison of Accuracy

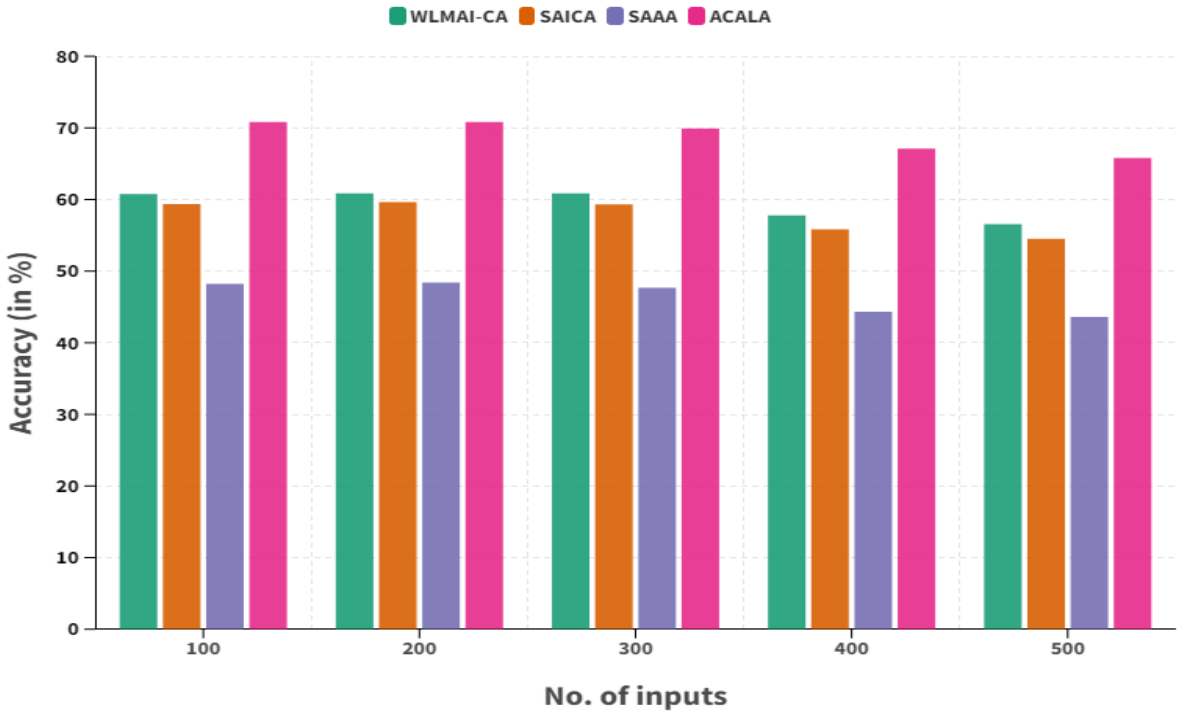

Fig.4 Comparison of Accuracy

4.3. Scalability: These AI-driven cloud architectures should be able to facilitate a new operator-developer workflow and handle diverse workloads while scaling up infinitely in their performance capabilities. For this one must utilize the algorithms which can learn and perform well concerning workload having four properties. Fig.5 shows the Comparison of Scalability

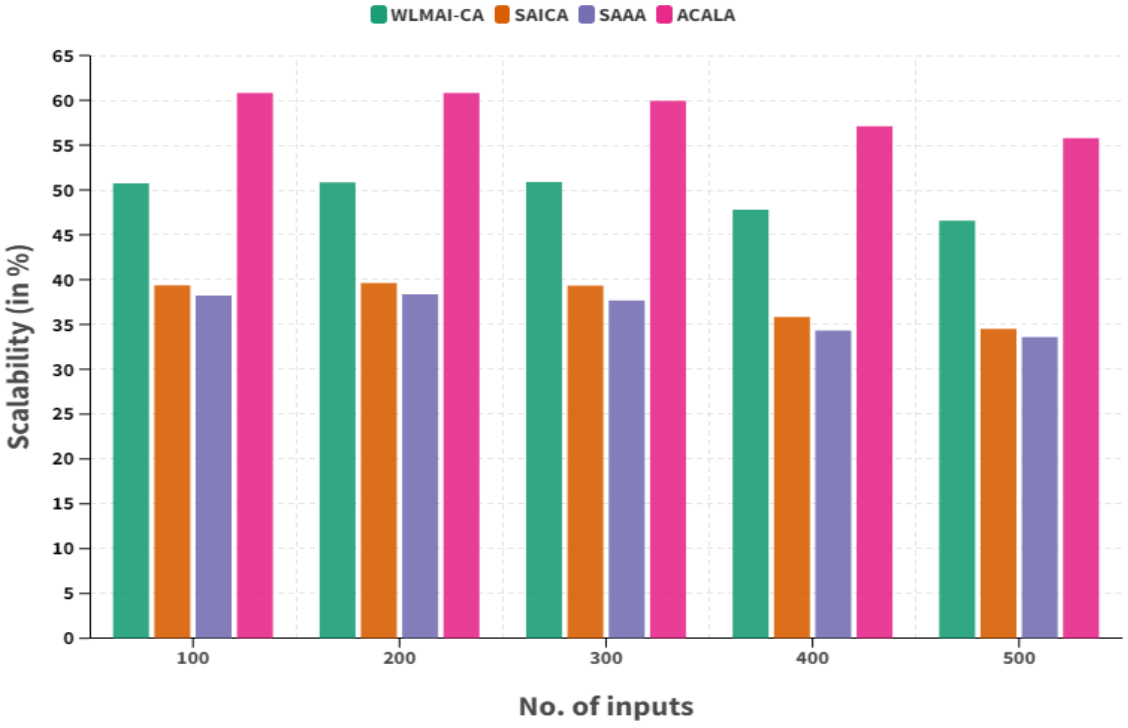

Fig.5 Comparison of Scalability

4.4. Resource Utilization: In AI-driven cloud architectures, cost-effective and sustainable operation is crucially dependent on the efficient use of resources. So, the algorithms will have to figure out how we get all that done without overburdening any one system resource until exhausted. Efficient use of computing, network bandwidth, and storage. Fig.6 shows the Comparison of Resource Utilization

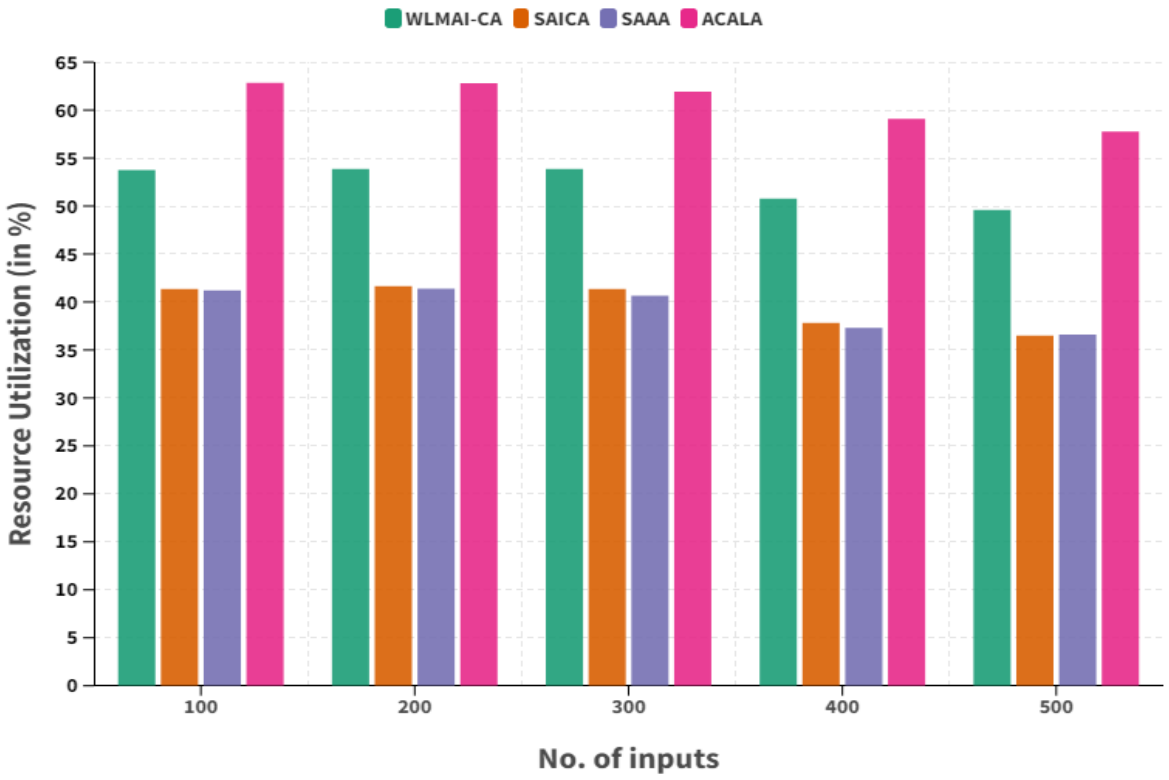

Fig.6 Comparison of Resource Utilization

5. Conclusion

Thereby we have concluded that AI-driven cloud architectures have revolutionized workload management in multiple sectors. The comparative algorithmic integrating form may set the workload management flow of an aggregator right, which will increase efficiency and productivity enough to cut down on cost without any performance decrease. AI algorithms like Predictive analytics and Machine learning which are independent of Human thought provide Automation in workload allocation, resource optimisation & decision-making. All in all, it comes down to making life easier for human managers helping them better focus their time and efforts on more strategic business tasks. In the future, as AI technology evolves even further, workload management in cloud architectures is expected to be faster than ever before revolutionizing organizational processing.

References

1. Liang, Q., Hanafy, W. A., Ali-Eldin, A., & Shenoy, P. (2023). Model-driven cluster resource management for ai workloads in edge clouds. ACM Transactions on Autonomous and Adaptive Systems, 18(1), 1-26.

2. Priyadarshini, S., Sawant, T. N., Bhimrao Yadav, G., Premalatha, J., & Pawar, S. R. (2024). Enhancing security and scalability by AI/ML workload optimization in the cloud. Cluster Computing, 1-15.

3. Bidollahkhani, M., Sharma, A. K., & Kunkel, J. M. (2024, July). HOSHMAND: Accelerated AI-Driven Scheduler Emulating Conventional Task Distribution Techniques for Cloud Workloads. In 2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC) (pp. 2313-2320). IEEE.

4. Tuli, S., Mirhakimi, F., Pallewatta, S., Zawad, S., Casale, G., Javadi, B., ... & Jennings, N. R. (2023). AI augmented Edge and Fog computing: Trends and challenges. Journal of Network and Computer Applications, 216, 103648.

5. Seo, C., Yoo, D., & Lee, Y. (2024). Empowering Sustainable Industrial and Service Systems through AI-Enhanced Cloud Resource Optimization. Sustainability, 16(12), 5095.

6. Singh, V., & Yadav, N. (2023). Optimizing Resource Allocation in Containerized Environments with AI-driven Performance Engineering. International Journal of Research Radicals in Multidisciplinary Fields, ISSN: 2960-043X, 2(2), 58-69.

7. Alsadie, D. (2024). A Comprehensive Review of AI Techniques for Resource Management in Fog Computing: Trends, Challenges and Future Directions. IEEE Access.

8. MURTHY, P., & BOBBA, S. (2021). AI-Powered Predictive Scaling in Cloud Computing: Enhancing Efficiency through Real-Time Workload Forecasting.

9. Ghelani, D. (2024). Optimizing Resource Allocation: Artificial Intelligence Techniques for Dynamic Task Scheduling in Cloud Computing Environments. International Journal of Advanced Engineering Technologies and Innovations, 1(3), 132-156.

10. Singh, A., & Aggarwal, A. (2023). 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, 3(2), 24-52.

11. Abdi, A., & Zeebaree, S. R. (2024). Embracing Distributed Systems for Efficient Cloud Resource Management: A Review of Techniques and Methodologies. Indonesian Journal of Computer Science, 13(2).

12. Firouzi, F., Farahani, B., & Marinšek, A. (2022). The convergence and interplay of edge, fog, and cloud in the AI-driven Internet of Things (IoT). Information Systems, 107, 101840.

13. Khan, M. A., & Walia, R. (2024, March). Intelligent Data Management in Cloud Using AI. In 2024 3rd International Conference for Innovation in Technology (INOCON) (pp. 1-6). IEEE.

14. Kommisetty, P. D. N. K. (2022). Leading the Future: Big Data Solutions, Cloud Migration, and AI-Driven Decision-Making in Modern Enterprises. Educational Administration: Theory and Practice, 28(03), 352-364.

15. Tatineni, S., & Chakilam, N. V. (2024). Integrating Artificial Intelligence with DevOps for Intelligent Infrastructure Management: Optimizing Resource Allocation and Performance in Cloud-Native Applications. Journal of Bioinformatics and Artificial Intelligence, 4(1), 109-142.

16. Alnafessah, A. (2022). Artificial intelligence driven anomaly detection for big data systems (Doctoral dissertation, Imperial College London).

17. Giagkos, D., Tzenetopoulos, A., Masouros, D., Xydis, S., Catthoor, F., & Soudris, D. (2024). AI-Driven QoS-Aware Scheduling for Serverless Video Analytics at the Edge. Information, 15(8), 480.

18. Kokkonen, H., Lovén, L., Motlagh, N. H., Kumar, A., Partala, J., Nguyen, T., ... & Riekki, J. (2022). Autonomy and intelligence in the computing continuum: Challenges, enablers, and future directions for orchestration. arXiv preprint arXiv:2205.01423.

19. Iriogbe, E. I. (2021). Reducing the Cloud Overhead and Latency for Artificial Intelligence Applications Using Hybrid Computing (Doctoral dissertation, Dublin, National College of Ireland).

20. Mills, N., Issadeen, Z., Matharaarachchi, A., Bandaragoda, T., De Silva, D., Jennings, A., & Manic, M. (2024). A cloud-based architecture for explainable Big Data analytics using self-structuring Artificial Intelligence. Discover Artificial Intelligence, 4(1), 33.

21. Daruvuri, R., Patibandla, K., & Mannem, P. (2024). Leveraging unsupervised learning for workload balancing and resource utilization in cloud architectures. International Research Journal of Modernization in Engineering Technology and Science, 6 (10), pp 1-8.