ISSN 2349-0292
Impact Factor 3.802
GLOBAL JOURNAL OF ADVANCED ENGINEERING TECHNOLOGIES AND SCIENCES

AN EFFICIENT SCHEDULING ALGORITHM FOR CLOUD SERVICE REQUESTS

Ayush Acharya*1, Arpit Solanki*2
*[1]Student, Dr. A.P.J. Abdul Kalam University, Indore, M.P., India
*[2]Assistant Professor, Dr. A.P.J. Abdul Kalam University, Indore, M.P., India

ABSTRACT

Cloud computing provides scalable and cost-effective computing services, but efficient scheduling of service requests remains a major challenge due to dynamic and heterogeneous workloads. Inefficient scheduling leads to increased execution time, poor resource utilization, and higher service costs.

This thesis proposes a Priority-based Batch Scheduling algorithm to reduce overall service utilization time in cloud computing environments. Service requests are classified into three priority levels based on Service Level Agreement (SLA) parameters. High-priority requests are executed immediately using FCFS scheduling, while medium- and low-priority requests are grouped into batches based on similar resource requirements to minimize communication and initialization delays.

Experimental results demonstrate that the proposed approach outperforms existing priority-based scheduling algorithms in terms of execution time, with performance improvements becoming more significant as the number of service requests increases. The proposed algorithm effectively enhances scheduling efficiency, reduces execution cost, and improves overall cloud system performance.

KEYWORDS: Cloud Computing, SLA, Service Request, Scheduling

INTRODUCTION

The term "cloud computing" describes a model in which consumers get computer capabilities—like processing power, storage, networks, and software—as services as opposed to locally held resources. This method relieves end users of the burden of managing hardware or system configurations as the cloud service provider maintains and controls the platforms and infrastructure. The National Institute of Standards and Technology (NIST) provides the most commonly accepted and authoritative definition of cloud computing.

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.- U.S. National Institute of Standards and Technology (NIST)".[14]

Services in the cloud environment are often divided into three main models:

(i) **Software as a Service (SaaS):** This business model uses the internet to distribute software programs. The program may be accessed by users without having to bother about infrastructure administration, installation, or maintenance.

(ii) **Platform as a Service (PaaS):** PaaS gives developers a platform to create, test, and implement applications. Without requiring users to manage underlying hardware or operating systems, it provides tools, libraries, and development environments.

(iii) **Infrastructure as a Service (IaaS):** This service model provides basic computer resources including networking, storage, and servers. With more flexibility and scalability, users may configure and administer these resources in a manner akin to that of conventional on-premises infrastructure.

In recent years, cloud computing has become increasingly popular in both the IT sector and educational institutions, because consumers only pay for the resources they really use under this paradigm, cloud services are very scalable and reasonably priced. As the demand for cloud services increases, service providers frequently get a high number of diverse and dynamic user requests. However, users have no knowledge of the server's load or resource availability. In such cases, handling and scheduling these requests is a crucial task. When the number of service requests grows, determining which one to process first becomes difficult, demanding clever scheduling solutions.

To overcome this issue, scheduling techniques are critical in cloud computing systems. Scheduling governs how resources are assigned to user requests in order to maximize performance, shorten wait times, and increase overall system efficiency. Without an appropriate scheduling mechanism, cloud systems may have resource

underutilization, high response times, cost inefficiencies, and low user satisfaction. As a result, effective scheduling is critical for reaching peak performance in cloud environments and managing the increasing number of applications and services that rely on them. In this manner, this paper proposes and implements an effective scheduling algorithm for cloud computing environments with the goal of lowering the overall cost of service utilization in the terms of execution time.

RELATED STUDY

This section outlines, some noteworthy contributions has been reported

P.Lahande et al. (2022) compared FCFS and SJF resource scheduling algorithms in a cloud environment using the WorkflowSim simulator and the Alibaba task event dataset. Their results show that SJF consistently outperforms FCFS in terms of average start time, completion time, waiting time, turnaround time, and cost. A. Kadhem et al. (2021), this paper examines job scheduling algorithms in cloud computing and discusses how they might improve system performance, resource utilisation, load balancing, and quality of service. It first covers the principles of cloud computing, service models (SaaS, PaaS, and IaaS), deployment types, and cloud architecture, before focussing on task scheduling as a critical difficulty in virtual resource management. The study analyses common scheduling algorithms—FCFS, SJF, Round Robin, Priority, Min-Min, and Max-Min—using examples and Gantt charts to examine waiting time, turnaround time, throughput, and CPU utilisation. The results demonstrate that Shortest Job First (SJF) is the most efficient fundamental algorithm in terms of average waiting and turnaround time, albeit it may induce starvation for long tasks.The research indicates that no single method is suitable for all cases, and the decision is determined by system needs such as delay sensitivity, fairness, and workload type. D.Daniel et al. (2011), The novel cloud scheduling scheme uses SLA (Service Level Agreement) along with trust monitor to provide a faster scheduling of the over flooding user request with secure processing of the request. The SLA criteria are used to ensure cloud security, and a third-party trust monitor keeps an eye on every service operations. Any unauthorised activity is deemed an intrusion, reported to the scheduler, and a notification is delivered to the user.

provider, and additional fines are imposed. M.Paul et al. (2011), In order to assess the complete collection of tasks in the task queue and determine the minimum completion time of all tasks, this paper employed credit-based scheduling decisions. The cost matrix in this case has been produced as a task's equitable propensity to be assigned to a resource. S.Kailsam et al. (2010), One method that adds a few characteristics to the scheduling methodology is optimising the Service Level Agreement (SLA). Here, the authors examined the relative performance of several algorithms in terms of speed, completion time, and use of resources. Z.Lee1 et al. (2011),The issue of service request scheduling in cloud computing systems was discussed by the author. We examine the resource suppliers, service providers, and consumers that make up the three-tier cloud system. In this case, the scheduling tactics for service requests should meet the goals of both customers and service providers. Jeevithra.R et al. (2018) proposed a task-aware priority-based scheduling algorithm for cloud computing to improve execution, waiting, and response times. The approach assigns three priority levels to tasks using statistical X-bar chart limits based on task characteristics. Simulation results using CloudSim show that the proposed method outperforms existing scheduling techniques in terms of efficiency and resource utilization, making it suitable for workflow-based cloud applications. A systematic review of cloud computing work scheduling techniques was published by S. Krishna et al. (2023), with an emphasis on machine learning, metaheuristic, and nature-inspired methods. The paper finds trust and fault tolerance as significant research gaps after analysing important performance metrics as makespan, cost, energy usage, response time, and resource utilisation. A. Behera et al. (2022) highlighted the importance of effective scheduling in lowering execution costs and enhancing resource utilisation. From the viewpoints of both users and providers, the article addresses cloud service models, scheduling issues, and QoS needs. It examines current energy-efficient and cost-effective scheduling strategies and comes to the conclusion that efficient resource scheduling is essential for cloud environments' cost control and performance optimisation.

PROBLEM STATEMENT

Cloud computing environments handle a large number of diverse and dynamic service requests submitted by users who have no visibility into the current workload or availability of cloud resources. As the volume of incoming requests increases, determining the order of execution and allocating appropriate resources become complex tasks. Existing cloud scheduling algorithms often fail to efficiently manage unpredictable workloads, leading to increased service execution time, poor resource utilization, higher operational costs, and degradation in overall quality of service. Therefore, there is a need for an effective cloud scheduling mechanism that can intelligently prioritize service requests and optimally allocate resources. The problem addressed in this research is the design

ISSN 2349-0292
Impact Factor 3.802

and development of an improved scheduling algorithm that minimizes service execution and waiting time while maximizing resource utilization and overall system performance in cloud computing environments.

PROPOSED WORKING MODEL

Batch processing is a novel concept in cloud computing. It involves grouping of similar types of jobs and processing them by a cloud service provider. This grouping approach reduces communication time or service initialization time between the cloud broker and the resources to reduce overall service time. The new technique requires only minor changes to the broker policy module and class files, with no requirement for hardware modifications.

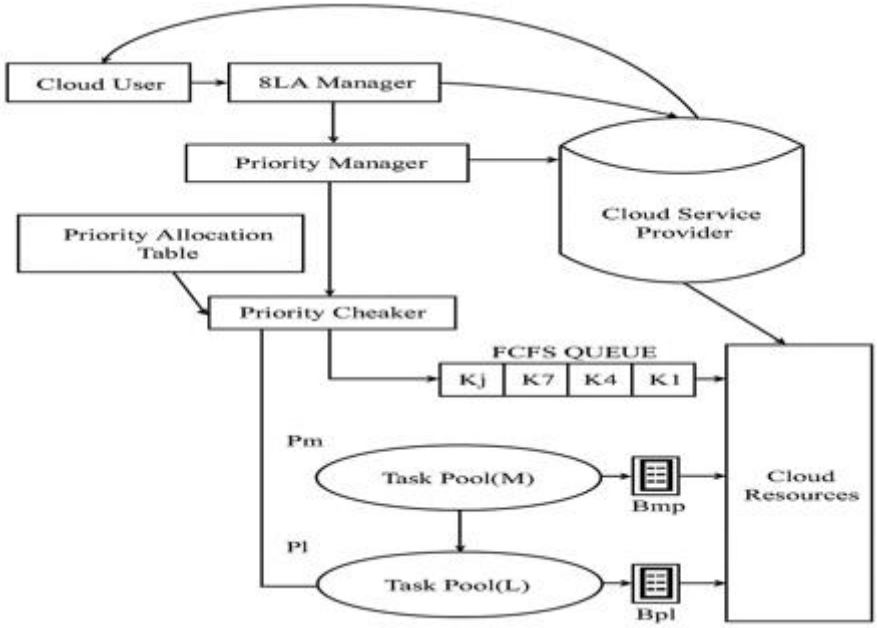

Fig : 1 Proposed Working Model

RESULT ANALYSIS

In this section, we analyze the performance of the proposed algorithm. A comparison with existing priority based scheduling algorithm has also been provided. The implemented algorithms were also introduced in the previous chapter. In this chapter, the experimental results are discussed.

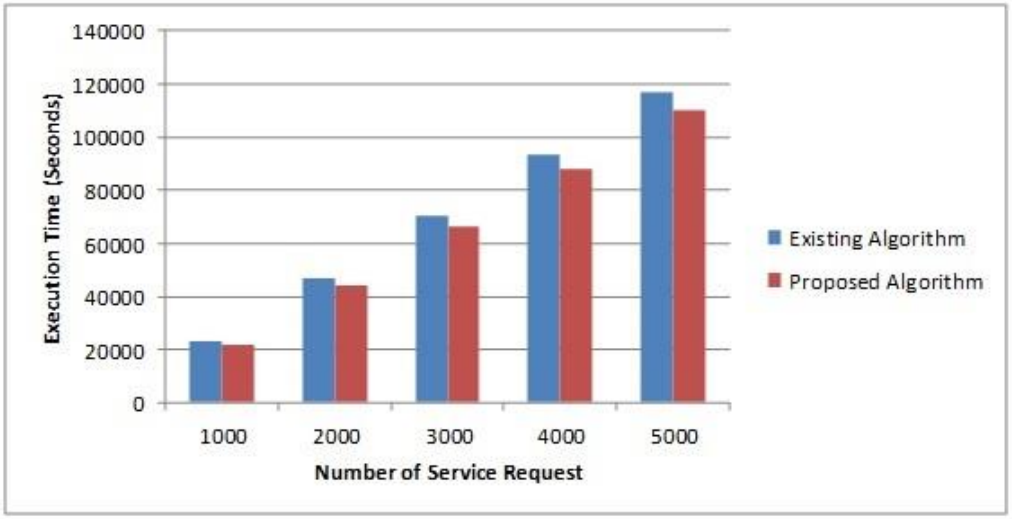

Fig: 2 Execution Time

As per the experimental result, comparison clearly shows that the proposed algorithm performs better in the terms of Execution time. The following Graph shows the improvement of proposed scheduling algorithm over the

existing priority based scheduling in cloud computing environment, with heterogeneous service requests. This demonstrates that the proposed approach taken less execution time.

CONCLUSION AND FUTURE SCOPE

Cloud computing has changed the way computational resources and services are supplied, providing greater flexibility, scalability, and cost-effectiveness. Effective scheduling of service requests is critical for optimizing resource utilization, reducing execution time, and lowering overall service costs. This study proposed a Priority-based Batch Scheduling algorithm that combines priority and batch processing to overcome the limitations associated with traditional priority-based scheduling.

The proposed working model introduces a structured mechanism where service requests are categorized into three priority levels according to SLA requirements. High-priority requests are executed immediately using FCFS scheduling, while medium- and low-priority requests are grouped into batches based on their resource requirements. This strategy effectively minimizes redundant communication between cloud brokers and data centers, which is a major source of delay in traditional scheduling techniques. Although batch formation introduces a small grouping time overhead, experimental results show that this overhead is negligible compared to the reduction achieved in communication delay, especially under heavy workloads. The results clearly demonstrate that the proposed scheduling algorithm consistently outperforms the existing priority-based scheduling method in terms of execution time. As the number of service requests increases, the performance improvement becomes more pronounced, highlighting the scalability and effectiveness of the proposed approach in handling large volumes of heterogeneous service requests. The proposed scheduling algorithm lays a solid platform for future study and development in cloud computing settings. Optimized computing is becoming increasingly significant for both the IT sector and educational institutions, particularly with the pay-as-you-go cloud service model. Future research could improve the suggested algorithm by including priority inversion handling to increase fairness and responsiveness. In addition, the scheduling can be improved by using machine learning algorithms that forecast service request patterns and dynamically alter batch grouping and priority levels. Energy-efficient scheduling can also help to reduce operational costs and environmental effect. Additional extensions can enable real-time and latency-sensitive applications, as well as scalability in hybrid and multi-cloud systems. Incorporating additional QoS criteria such as reliability, availability, throughput, and fault tolerance would strengthen and broaden the scheduling approach's application in real-world cloud environments.

REFERENCES

[1] P. Lahande, P. Kaveri, “Implementing FCFS and SJF for finding the need of Reinforcement ITM Web of Conferences 50, 01004 (2022) ICAECT 2022
[2] A. Kadhem, H. Aziz, T.Y.Gasim, H.Ismaeel “Task Scheduling Algorithms in Cloud Computing”, Easy Chair Print, Islamic Azad University South Tehran Branches
[3] Jeevithra.R, Karthikeyan.T “Priority Based Scheduling In Cloud Computing”, International Journal of Computer Techniques-Volume 5, Issue 4 July - Aug 2018
[4] M.S.R. Krishna, S.Mangalampalli “A Systematic Review on Various Task Scheduling Algorithms in Cloud Computing”, EAI Endorsed Transactions on Internet of Things, Volume 10,2024.
[5] A. Behera, B.K. Pattanayak “A Study on Resource Scheduling Techniques in Cloud Computing Environment”, International Journal of Advances in Engineering and Management (IJAEM)
[6] D.Daniel , S.P.Jeno Lovesum,“A Novel Approach for Scheduling Service Request in Cloud with Trust Monitor”, Proceedings of 2011 International Conference on Signal Processing, Communication, Computing and Networking Technologies (ICSCCN 2011).
[7] Zhongyuan Lee1, Ying Wang1, Wen Zhou2,“A dynamic priority scheduling algorithm on service request scheduling in cloud computing”, 2011 International Conference on Electronic & Mechanical Engineering and Information Technology..
[8] Mousumi Paul, Goutam Sanyal,“ Task-Scheduling in Cloud Computing using Credit Based Assignment Problem”, International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 10 October 2011..
[9] [9] Fang Dong, Junzhou Luo, Lisha Gao and Liang Ge, “A Grid Task Scheduling Algorithm Based on QoS Priority Grouping”, Fifth IEEE International Conference on Grid and Cooperative Computing (GCC’06).
[10] Amazon Elastic Compute Cloud, “Amazon EC2”http://aws.amazon.com/ec2/.

[11] R. Buyya, C.S. Yeo, and S. Venugopal, " Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities", In High Performance Computing and Communications, 2008. HPCC08. 10th IEEE International Conference on, pages 513. IEEE, 2008.
[12] DC Plummer, TJ Bittman, T. Austin, D. Clearley, and DM Smith "Cloud computing: Defining and describing and emerging phenomenon", Gartner, Inc. Retrieved September, 25:2008, 2008.
[13] J. Staten, "Is cloud computing ready for the enterprise?",Forrester Research, March,7, 2008.
[14] P. Mell and T. Grance, "The NIST definition of cloud computing", National Institute of Standards and Technology, 2009.
[15] "Think Grid Business IT on Demand", http://www.thinkgrid.co.uk/.
[16] Michael Behrendt, Bernard Glasner, Petra Kopp, at. el.,"IBM Cloud Computing Reference Architecture", Date. Feb 2011
[17] S.M. Shatz, J. P. Wang, and M. Goto,"Task allocation for maximizing reliability of distributed computing system", IEEE trans. computers, vol, 41, no. 9, pp. 1156 -1168,1992.
[18] "Shirley Radack NIST.gov publication"http://csrc.nist.gov/publications/nistbul/march-2012 itlbulletin.pdf
[19] Dukee, D. "Why cloud computing will never be free.", Commun, ACM 53, 5(May.2010), 62-69
[20] Hai Zhong, Kun Tao and Xueji Zhang, "An Approach to Optimized Resource Scheduling Algorithm for Open-source Cloud Systems", The Fifth Annual ChinaGrid Conference.
[21] Rodrigo N. Calheiros, Rajiv Ranjan, Anton Beloglazov, Cesar A.F.De Rose and Rajkumar Buyya, "CloudSim: A Toolkit for Modeling and Simulation of Cloud Computing Environments and Evaluation of Resource Provisioning Algorithms".
[22] Sriram Kailsam, Nathan Gnanasambam, Jananiram Dharanipragad and Naveen Sharma, "Optimizing Service Level Agreements for Automatic Cloud Bursting Schedulers", 39th International Conference on Parallel Processing Workshops,2010.
[23] [23] Introduction to cloud computing,"Cloud computing - Wikipedia, the free encyclopedia htm",http://en.wikipedia.org/wiki/Cloud computing.
[24] https://www.salesforce.com/in/blog/what-is-iaas-paas-saas
[25] https://www.geeksforgeeks.org/computer-science-fundamentals/cloud-deployment-models
essing Using Matlab and Wavelets, Infinity Science Press LLC, 2007.