JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Optimizing Resource Allocation and Workload Management in Multi-Cloud Environments: A Review Study

Amit Kumar, Ph.D. (Part Time)
Dept. of Electronic and Communication Engineering
Sarala Birla University, Namkum, Ranchi

Abstract: Optimizing resource allocation and workload management in multi-cloud environments is essential for maximizing efficiency and cost-effectiveness. This paper investigates the challenges and strategies involved in managing resources across diverse cloud platforms. It explores the evolution of cloud computing, emphasizing the significance of effective resource management strategies in overcoming interoperability, security, and cost management challenges. Various resource allocation techniques and workload management strategies are discussed, along with their implications for multi-cloud environments. By providing a comprehensive overview, this paper aims to offer insights and recommendations to help organizations navigate the complexities of multi-cloud resource management effectively.

Keywords: multi-cloud environments, resource allocation, workload management, optimization.

I. Introduction

In today's digital landscape, the adoption of multi-cloud environments has become increasingly prevalent as organizations seek to leverage the benefits of diverse cloud platforms. However, efficiently managing resources and workloads across these multiple cloud infrastructures presents a complex challenge. This review paper explores the critical aspects of optimizing resource allocation and workload management within multi-cloud environments. By examining the evolution of cloud computing and the emergence of multi-cloud architectures, the paper sets the stage for understanding the significance of effective resource management strategies. The introduction highlights the importance of this topic in overcoming challenges such as interoperability issues, security concerns, and cost management complexities. Furthermore, it outlines the scope of the paper, including a discussion of various resource allocation techniques, workload management strategies, challenges faced in multi-cloud environments, and opportunities for future advancements. By providing a comprehensive overview of the subject matter, this paper aims to offer insights and recommendations to help organizations navigate the complexities of multi-cloud resource management effectively. [1]

A. Multi-Cloud Environments

Multi-cloud environments refer to the use of multiple cloud computing services or platforms from different providers to fulfil an organization's computing needs. These environments enable businesses to leverage the strengths of various cloud services while mitigating the risks associated with vendor lock-in and single-point failures. By distributing workloads across multiple clouds, organizations can achieve better scalability, reliability, and flexibility. Multi-cloud strategies often involve a combination of public, private, and hybrid cloud deployments, allowing organizations to optimize resource usage, enhance performance, and reduce costs. However, managing resources and workloads across disparate cloud platforms presents unique challenges that require careful consideration and strategic planning. [2]

B. Importance of Resource Allocation and Workload Management

Effective resource allocation and workload management are paramount in multi-cloud environments due to their direct impact on performance, cost efficiency, and overall operational effectiveness. In the dynamic landscape of cloud computing, where resources are shared among multiple users and across different cloud platforms, optimizing resource allocation

ensures that computing resources such as processing power, storage, and network bandwidth are allocated efficiently to meet the demands of varying workloads. Proper resource allocation enables organizations to make the most of their cloud investments by ensuring that resources are neither underutilized nor overprovisioned. Underutilization leads to wasted resources and increased costs, while overprovisioning results in unnecessary expenses and inefficient resource utilization. By implementing effective resource allocation strategies, organizations can achieve better resource utilization rates, thereby optimizing their cloud spending and maximizing ROI. Moreover, efficient workload management plays a crucial role in maintaining the performance and responsiveness of applications and services hosted in multi-cloud environments. Workload management involves tasks such as workload distribution, task scheduling, and prioritization, ensuring that workloads are processed timely and efficiently across different cloud platforms. Proper workload management helps prevent resource bottlenecks, reduces latency, and ensures a consistent user experience. [1-3] In multi-cloud environments, where workloads may span across multiple cloud providers and data centres, effective workload management becomes even more critical to ensure seamless operation and avoid vendor lock-in. By intelligently distributing workloads and managing dependencies, organizations can enhance fault tolerance, resilience, and scalability, ultimately improving their overall business agility and competitiveness in the digital era. [4]

II. Background

Alyas et al. (2023) mentioned that cloud computing made dynamic resource provisioning more accessible. They emphasized the importance of monitoring a functioning service, stating that changes were made when particular criteria were surpassed. Their research explored the decentralized multi-cloud environment for allocating resources and ensuring the Quality of Service (QoS), estimating the required resources, and modifying allotted resources depending on workload and parallelism due to resources. They found that resource allocation was a complex challenge due to the versatile service providers and resource providers. They stressed the need for a cooperation strategy for sustainable quality of service, emphasizing the objective of achieving a coherent and rational resource allocation to attain the quality of service. They also included identifying critical parameters to develop a resource allocation mechanism. They proposed a framework based on the specified parameters to formulate a resource allocation process in a decentralized multi-cloud environment. They further divided three main parameters of the proposed framework, namely data accessibility, optimization, and collaboration, into subsets for resource allocation and long-term service quality using an optimization technique. The CloudSim simulator was used to validate the suggested framework through several experiments aimed at finding the best configurations suited for enhancing collaboration and resource allocation to achieve sustained QoS. The results supported the suggested structure for a decentralized multi-cloud environment and the parameters that had been determined.

Sangeetha et al. (2022) discussed the wide increase in cloud traffic due to the enormous increase of media content in recent decades. They highlighted the popularity of cloud due to its ease of use and flexible model but noted its historical issues with poor resource management and minimal extendibility of service portfolio. However, they mentioned recent advancements in effectively managing services and making their discovery further possible. They emphasized the need for wide operable systems to handle larger amounts of multimedia content in a standalone cloud effectively. They presented a resource allocation framework using a gray wolf optimization (GWO) architecture to effectively learn the operation of resource allocation in an optimal manner. They described how the cloud at times communicates with each other based on the resource allocated by the deep neural network and then shares resources, forming the multi-cloud computing. They explained how the deep neural network acts as a model for controlling routing capabilities based on input data rate and available storage space in the multi-clouds, reducing the delay in processing and storage of data to ensure flexible operations across the cloud. They divided the entire operation into two modules: the first module operated on data processing and routing operations, while the second module acted as a control plane using the deep neural network to ensure optimal allocation of resources based on data obtained and processed in the first module. They highlighted how these two models ensured better delivery of data to the cloud with proper allocation and storage of resources in the multi-cloud environment. They conducted simulations using the Java (NetBeans) platform and evaluated them further using the CloudSim toolkit. The results were experimented on various performance metrics that included time delay and cost of resource allocation on multi-cloud.

Rajeshwari et al. (2022) explained how a single cloud provider could manage their cloud on their own individually to deliver services to users based on the need and demand of the user anywhere, anytime. They discussed how multi-cloud technology had been proposed as a revolutionary solution in recent years, where a user uses more than one cloud platform and each one of them delivers a specific application or service, allowing for the distribution of total workload across multiple clouds. They described the challenges of collaboration among cloud service providers in the heterogeneous multi-cloud environment and workload balancing among multiple clouds, highlighting the benefits such as load balancing, reduced waiting time, improved response time of requests, avoidance of service level agreement violations, effective utilization of resources, and optimization of power usage at each cloud. They presented the architecture and functionality of multi-cloud,

managing multi-cloud environment such as balanced distribution of load among multiple clouds, power management in an optimized way at each cloud, monitoring of service level agreement, and satisfying service level agreement in the first part of their book chapter. They discussed the current challenges in multi-cloud like balanced distribution of workload among multiple cloud, application and service deployment in a heterogeneous multi-cloud environment, use of resources and services from different cloud service providers, security in data storage, large scale data processing, and research directions toward the multi-cloud in the second part. Finally, they elaborated on the simulators available for multi-cloud, applicability of simulators, and best simulators to solve research challenges in multi-cloud management, providing insights into choosing the best and suitable simulator for solving different research issues.

Manekar & Pradeepini (2021) stated that cloud computing was most powerful and demanding for businesses in the last decade. They mentioned the challenges faced in accommodating and processing data efficiently, emphasizing the need for further research toward big data processing in multi-cloud infrastructure. They proposed an effective deadline-aware resource management scheme through novel algorithms, namely job tracking, resource estimation, and resource allocation. They discussed two algorithms in detail and conducted an experiment in a multi-cloud environment, checking job track algorithms firstly and job estimation algorithms lastly. They highlighted the utilization of multiple cloud service providers as a promising solution for an affordable class of services and QoS.

Chen et al. (2020) mentioned that emerging multi-cloud environments (MCEs) empowered the execution of large-scale scientific workflows (SWs) with sufficient resource provisioning. They addressed the challenges faced in SW scheduling in MCEs due to complex task dependencies in SWs and various cost-performance of cloud resources. They proposed an Online Workflow Scheduling algorithm based on Adaptive resource Allocation and Consolidation (OWS-A2C) to address these challenges. They explained the execution of deadline reassignment for SW tasks based on the execution performance of instance resources in OWS-A2C, which enhanced resource utilization from a local perspective. They described the allocation and consolidation of execution instances according to the performance requirements of multiple SWs to improve resource utilization and reduce total costs from a global perspective. They outlined how SW tasks were dynamically scheduled to execution instances with the earliest-deadline-first (EDF) discipline and completed before their sub-deadlines in OWS-A2C. They conducted extensive simulation experiments to demonstrate the effectiveness of OWS-A2C on SW scheduling in MCEs, showing higher resource utilization and lower execution costs under deadline constraints compared to three baseline scheduling methods.

Sadashiv & Kumar (2018) highlighted the widespread adoption of cloud computing by small, medium, and large business organizations to host interactive web-based applications due to their unlimited services compared with the classical computing approach. They discussed the challenge faced by cloud service providers in providing uninterrupted service at an economical price with efficient utilization of resources, especially in serving users spread across the globe. They presented a resource management approach for deploying three-tier applications over a broker-based multi-cloud environment, considering strategies for quick cloud site selection, dynamic resource adaptation, and two-level load balancing with high availability. They conducted experiments on an extended CloudSim simulator using realistic session workloads synthesized based on different statistical distributions. They evaluated the approach's performance, revealing that the strategies led to improved resource utilization, throughput, and compliance with SLA even under varying workload scenarios.

Alsarhan et al. (2017) discussed the cloud market where various resources such as CPUs, memory, and storage in the form of Virtual Machine (VM) instances could be provisioned and then leased to clients with QoS guarantees. They proposed a novel Service Level Agreement (SLA) framework for cloud computing, utilizing reinforcement learning (RL) to derive a VM hiring policy that could adapt to changes in the system to guarantee the QoS for all client classes. They integrated computing resources adaptation with service admission control based on the RL model to enhance the CP's profit and avoid SLA violation. They conducted numerical analysis to stress the ability of their approach to avoid SLA violation while maximizing the CP's profit under varying cloud environment conditions.

Table 1. Comparative Reviews

Author(s) (Year)	Area of Research	Methodology	Conclusion
Alyas et al. (2023)	Decentralized Multi-Cloud Resource Allocation	Proposal of a framework for resource allocation in a decentralized multi-cloud environment based on specified parameters.	Supported structure for decentralized multi-cloud environment and identified parameters for resource allocation mechanism.
Sangeetha et al. (2022)	Resource Allocation in Multi-Cloud Environment	Design of a resource allocation framework using GWO and deep neural network, simulation, and evaluation.	Validation of proposed framework through simulation, supporting efficient resource allocation and improved service delivery in multi-cloud environment.
Rajeshwari et al. (2022)	Multi-Cloud Architecture and Management	Presentation of multi-cloud architecture, challenges, and strategies for load balancing and resource management.	Discussion on the effectiveness of multi-cloud architecture, challenges, and strategies for improved performance and resource utilization.
Manekar & Pradeepini (2021)	Deadline-Aware Resource Management	Proposal of algorithms for resource management and experimentation in a multi-cloud environment.	Highlighted the potential of utilizing multiple cloud service providers for cost-effective services and quality of service improvements.
Chen et al. (2020)	Scientific Workflow Scheduling in MCEs	Proposal of OWS-A2C algorithm, simulation, and evaluation.	Demonstrated effectiveness of OWS-A2C algorithm in improving resource utilization and reducing execution costs in multi-cloud environments.
Sadashiv & Kumar (2018)	Resource Management in Broker-Based Multi-Cloud	Presentation of resource management strategies and experimentation in a multi-cloud environment.	Demonstrated improved resource utilization, throughput, and compliance with SLA in multi-cloud environments through proposed strategies.
Alsarhan et al. (2017)	Cloud Service Level Agreement (SLA) Framework	Proposal of SLA framework using RL, numerical analysis.	Demonstrated the ability of the proposed SLA framework to avoid SLA violations and maximize profit for cloud providers under varying conditions.

III. Resource Allocation Techniques

A. Static Allocation Strategies

In multi-cloud environments, resource allocation techniques play a crucial role in ensuring efficient utilization of computing resources across various cloud platforms. This section discusses two primary categories of resource allocation strategies: static allocation and dynamic allocation.

A. Static Allocation Strategies

Round-Robin Allocation: Round-robin allocation is a simple static allocation strategy that evenly distributes incoming workload requests across available cloud resources in a cyclical manner. This technique ensures fair resource utilization among multiple cloud instances or servers without considering their current load or capacity. Each subsequent request is directed to the next available resource in a circular sequence, ensuring that no resource is favoured over others. Round-robin allocation is easy to implement and suitable for scenarios where workload demands are relatively uniform across cloud resources.

Weighted Allocation: Weighted allocation is a static allocation strategy that assigns priorities or weights to different cloud resources based on predefined criteria such as resource capacity, performance metrics, or service level agreements (SLAs). Resources with higher weights receive a larger share of incoming workload requests, while those with lower weights handle fewer requests. This technique allows organizations to allocate resources based on their capabilities and requirements, ensuring that critical workloads receive preferential treatment over less critical ones. Weighted allocation provides flexibility in resource management and enables organizations to optimize resource allocation based on workload priorities and business objectives. Static allocation strategies like round-robin and weighted allocation offer simplicity and predictability in resource management, making them suitable for certain use cases in multi-cloud environments. However, they may not adapt well to dynamic workload fluctuations or optimize resource utilization in real-time. Dynamic allocation

strategies, discussed in the next section, address these limitations by adjusting resource allocation dynamically based on workload demand and system conditions. [5]

B. Dynamic Allocation Strategies

Dynamic allocation strategies in multi-cloud environments focus on optimizing resource allocation in real-time based on changing workload demands and system conditions. This section discusses two key dynamic allocation strategies: load balancing algorithms and predictive analytics for resource allocation.

Load Balancing Algorithms

Load balancing algorithms are dynamic allocation strategies that distribute incoming workload requests across available cloud resources in a way that minimizes resource contention and maximizes overall system performance. These algorithms take into account various factors such as current resource utilization, response times, and available capacity to make informed decisions about workload distribution. Common load balancing algorithms include Round Robin, Least Connection, Weighted Round Robin, and Least Response Time.

- ✓ Round Robin: Distributes incoming requests evenly across available resources in a cyclical manner.
- ✓ Least Connection: Routes requests to the server with the fewest active connections, aiming to distribute the load more evenly.
- ✓ Weighted Round Robin: Similar to Round Robin, but assigns weights to resources based on their capacities or performance metrics to handle varying workload demands.
- ✓ Least Response Time: Routes requests to the server with the shortest response time, aiming to minimize user latency and improve overall system performance.

Predictive Analytics for Resource Allocation

Predictive analytics leverages historical data, machine learning algorithms, and predictive models to forecast future workload demands and resource requirements in multi-cloud environments. By analyzing past usage patterns, workload characteristics, and system performance metrics, predictive analytics can anticipate future resource needs and dynamically adjust resource allocation accordingly. This proactive approach enables organizations to optimize resource utilization, prevent resource bottlenecks, and improve overall system efficiency. Dynamic allocation strategies like load balancing algorithms and predictive analytics enable organizations to adapt to changing workload dynamics and optimize resource allocation in real-time, thereby improving system performance, scalability, and cost-effectiveness in multi-cloud environments. [6]

C. Hybrid Allocation Approaches

Hybrid allocation approaches combine elements of both static and dynamic allocation strategies to optimize resource allocation in multi-cloud environments. These approaches leverage the strengths of both static and dynamic techniques to achieve better resource utilization, performance, and cost efficiency. This section discusses key hybrid allocation approaches commonly used in multi-cloud environments:

i. Rule-Based Hybrid Allocation: Rule-based hybrid allocation approaches define a set of rules or policies that dictate resource allocation decisions based on predefined criteria such as workload characteristics, priority levels, and system requirements. These rules combine static allocation principles with dynamic adjustments based on real-time conditions. For example, organizations may define rules to allocate resources statically during normal workload conditions but switch to dynamic load balancing during peak demand periods to ensure optimal performance and scalability.

ii. Capacity-On-Demand Hybrid Allocation: Capacity-on-demand hybrid allocation involves dynamically provisioning additional resources from public cloud providers or extending resources from private cloud environments to meet sudden spikes in workload demand. This approach combines the scalability and flexibility of public clouds with the security and control of private clouds, allowing organizations to scale their resources dynamically while maintaining control over sensitive data and critical workloads. Capacity-on-demand hybrid allocation enables organizations to optimize resource usage and reduce costs by only provisioning additional resources when needed, thereby avoiding overprovisioning.

iii. Auto-Scaling Hybrid Allocation: Auto-scaling hybrid allocation utilizes automated scaling mechanisms to adjust resource allocation dynamically based on workload fluctuations and performance metrics. This approach leverages both static provisioning and dynamic scaling to ensure optimal resource utilization while meeting performance

targets and cost constraints. Auto-scaling hybrid allocation allows organizations to automatically scale resources up or down in response to changing workload demands, minimizing manual intervention and optimizing resource usage across multi-cloud environments.[7]

IV. Workload Management Strategies

A. Task Scheduling Algorithms

Task scheduling algorithms are critical for efficient workload management in multi-cloud environments, ensuring that computing tasks are executed effectively across various cloud resources. This section explores three fundamental task scheduling algorithms commonly used in multi-cloud environments:

a. First Come First Serve (FCFS): FCFS is one of the simplest task scheduling algorithms, where tasks are executed in the order they arrive in the system. In a multi-cloud environment, FCFS schedules incoming tasks to available cloud resources based on their arrival time, without considering their execution time or resource requirements. While FCFS is easy to implement and ensures fairness in task execution, it may lead to inefficient resource utilization, especially if long-running tasks occupy resources while shorter tasks remain in the queue.

b. Shortest Job Next (SJN): SJN is a non-pre-emptive task scheduling algorithm that prioritizes tasks based on their execution time. In a multi-cloud environment, SJN schedules tasks to available resources based on their estimated execution time, with shorter tasks being executed before longer ones. This algorithm aims to minimize average waiting time and turnaround time, thus improving overall system efficiency. However, SJN may suffer from starvation issues, where long-running tasks are continuously postponed in Favour of shorter tasks, leading to delays in task completion.

c. Round Robin Scheduling: Round Robin scheduling is a pre-emptive task scheduling algorithm that allocates fixed time slices, known as time quantum, to each task in a cyclic manner. In a multi-cloud environment, Round Robin scheduling distributes available computing resources among multiple tasks by allowing each task to execute for a specified time quantum before switching to the next task in the queue. This algorithm ensures fair allocation of resources and prevents starvation by regularly rotating tasks, but it may lead to increased context switching overhead and may not be suitable for tasks with varying execution times. [6-8]

B. Machine Learning-based Workload Management

Machine learning-based workload management is a cutting-edge approach that harnesses the power of machine learning algorithms to optimize resource allocation and task scheduling in multi-cloud environments. This methodology relies on historical workload data, real-time performance metrics, and predictive analytics to make informed decisions and dynamically adjust resource allocation based on changing workload patterns. By leveraging machine learning models, organizations can predict future workload demands more accurately and allocate resources efficiently to meet these demands. The implementation of machine learning-based workload management involves several key steps. Firstly, relevant data including workload characteristics, resource usage, and system performance metrics are collected and pre-processed. Next, features are extracted or engineered from the data, and machine learning models such as regression, classification, clustering, or reinforcement learning algorithms are selected and trained using historical data. Trained models are then used to predict future workload patterns and optimize resource allocation and task scheduling decisions in real-time based on predicted demands and system constraints.

The adoption of machine learning-based workload management offers several benefits for organizations operating in multi-cloud environments. These benefits include improved resource utilization, enhanced system performance and responsiveness, and cost optimization by minimizing unnecessary resource usage and optimizing cloud spending. However, challenges such as ensuring data quality and availability, managing model complexity and scalability, and enabling real-time adaptation to changing workload patterns need to be addressed to realize the full potential of machine learning-based workload management in multi-cloud environments. [8]

VIII. Conclusion and Future work

The optimizing resource allocation and workload management in multi-cloud environments is crucial for organizations to achieve optimal performance, scalability, and cost efficiency. Through the exploration of diverse resource allocation techniques and workload management strategies, this paper has highlighted the complexity of managing resources across multiple cloud platforms. While static allocation strategies offer simplicity and predictability, dynamic allocation

approaches, including load balancing algorithms and predictive analytics, enable real-time optimization based on changing workload dynamics. Hybrid allocation approaches combine the strengths of static and dynamic strategies to achieve better resource utilization and performance. Furthermore, machine learning-based workload management presents promising opportunities for enhancing resource allocation efficiency and responsiveness. Future research should focus on addressing challenges such as data quality, model scalability, and real-time adaptation to unlock the full potential of multi-cloud resource management.

References

1. Alyas, T., Ghazal, T. M., Alfurhood, B. S., Issa, G. F., Thawabeh, O. A., & Abbas, Q. (2023). Optimizing Resource Allocation Framework for Multi-Cloud Environment. *Computers, Materials & Continua*, *75*(2).
2. Sangeetha, S. B., Sabitha, R., Dhiyanesh, B., Kiruthiga, G., Yuvaraj, N., & Raja, R. A. (2022). Resource management framework using deep neural networks in multi-cloud environment. *Operationalizing Multi-Cloud Environments: Technologies, Tools and Use Cases*, 89-104.
3. Rajeshwari, B. S., Dakshayini, M., & Guruprasad, H. S. (2022). Workload Balancing in a Multi-Cloud Environment: Challenges and Research Directions. *Operationalizing Multi-Cloud Environments: Technologies, Tools and Use Cases*, 129-144.
4. Manekar, A., & Pradeepini, G. (2021). Optimizing cost and maximizing profit for multi-cloud-based big data computing by deadline-aware optimize resource allocation. In *Recent Studies on Computational Intelligence: Doctoral Symposium on Computational Intelligence (DoSCI 2020)* (pp. 29-38). Springer Singapore.
5. Chen, Z., Lin, K., Lin, B., Chen, X., Zheng, X., & Rong, C. (2020). Adaptive resource allocation and consolidation for scientific workflow scheduling in multi-cloud environments. *IEEE Access*, *8*, 190173-190183.
6. Chen, Z., Lin, K., Lin, B., Chen, X., Zheng, X., & Rong, C. (2020). Adaptive resource allocation and consolidation for scientific workflow scheduling in multi-cloud environments. *IEEE Access*, *8*, 190173-190183.
7. Sadashiv, N., & Kumar, S. D. (2018). Broker-based resource management in dynamic multi-cloud environment. *International Journal of High-Performance Computing and Networking*, *12*(1), 94-109.
8. Alsarhan, A., Itradat, A., Al-Dubai, A. Y., Zomaya, A. Y., & Min, G. (2017). Adaptive resource allocation and provisioning in multi-service cloud environments. *IEEE Transactions on Parallel and Distributed Systems*, *29*(1), 31-42.