



International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

CLOUD COMPUTING RESOURCE ALLOCATION

MR. MANISH KUMAR SONI¹, PRINCEE KATYAR²

¹GUIDE NAME Bansal Institute of Engineering and Technology APJ ABDUL KALAM TECHNICAL UNIVERSITY AKTU

²Bansal Institute of Engineering and Technology APJ ABDUL KALAM TECHNICAL UNIVERSITY AKTU

INTRODUCTION

“Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction.”

The cloud computing paradigm offers cost savings because users lease the computing resources from a service provider when needed instead of owning them. Cloud computing enables dynamic sharing of the computing resources among the users and allows them to submit and execute applications of many kinds. Different applications, however, require distinct types of resources. Applications such as interactive databases or web-based services lease the cloud resources and usually occupy various resources to maintain a high quality of service (QoS) level for their users. A service level agreement (SLA) specifies the QoS to be provided to the user in terms of various performance parameters such as throughput, reliability, blocking probability and response time.

The enabling technology of cloud computing is virtualization. Virtualization is a technique for separating software applications from each other and the hardware resources. Virtual Machine (VM) also called instance refers to running of an individual copy of a particular user's application software or operating system in a virtual environment. Virtual systems feature multitenant capabilities through hypervisor. Hypervisor (also called a virtual machine monitor) is the virtualization software platform that allows simultaneous running of multiple instances on a server. Housing multiple virtual machines on a physical server utilize the physical server resources more efficiently. In a non-virtualized environment, most of the time servers are idle and it has been found that less than 10 percent of computing resources are used at any one time. These servers accrue maintenance and human resource costs along with

the costs of the energy required to power and cool the hardware. After virtualization, server utilization is found to rise as high as 80 percent. Therefore, virtualization is one of the operative ways to mitigate datacenter (DC) expenses. Moreover, virtualized datacenters accommodate pools of resources rather than isolated servers. The goal is to pool resources and serve demands from these resource pools.

One of the prime objectives of modern computing is to implement parallel computations on large distributed resources such as DCs. On demand allocation of resources in a cloud DC improves performance and reduces the deployment overheads significantly. However, efficient resource allocation requires accurate modeling of DCs. This is why more accurate modeling of jobs and resources that reflect real world scenarios, has attracted wide attention recently. Although the cloud computing technology is emerging and growing rapidly, due to its complexity it suffers from a lack of standards [2], detailed models and optimal resource allocation algorithms.

The main issue that makes cloud resource allocation complex is heterogeneity of resources and workloads. Based on the types of applications served by the cloud computing center, there is a vast diversity in resource demand profiles which make the incoming workflow heterogeneous. Besides heterogeneity of workloads, one of the essential characteristics of a cloud computing system is heterogeneity of server resources. As time goes by, DCs update their resources' configuration, processing capabilities, memory and storage spaces. They also introduce new platforms based on the new high performance servers along with older hardware, which makes the cloud platform heterogeneous.

Cloud workloads often have very large range of resource requirements, arrival rate and execution time. Jobs entering the cloud may demand different types of services. In the interactive applications, the users occupy cloud resources for a long period of time, whereas batch jobs require much shorter execution time. Hence, the execution time of the applications is notably divergent. In general, computing jobs such as web serving are more processing intensive, while database operations typically require high-memory support. Most of the jobs require parallel data analysis. This is the main reason for the recent development of MapReduce Programming model [3]. This model relies on parallel processing with a sequential functional approach. Job fragments are executed in parallel to speed up processing of the jobs. MapReduce has usually three phases: fan out, map and reduce. Applications such as Apache Hadoop [4] and platforms such as Pig [5] implement the MapReduce programming model. This model also applies to bag-of-tasks (BoTs) where a job consists of parallel and sequential tasks. The number of task executions in the mapping phase will be larger than the fan out and reduce phases. Hence, a job's load depends on the phase of execution.

As a result, high variety in resource and workload characteristics makes resource allocation in cloud computing centers very challenging.

Present day servers consume more power than those of a decade ago. Researchers believe that high performance servers impose more energy costs to the system. With the growing number of in-service servers, the worldwide expenditure on enterprise power usage and server cooling is estimated to be quite high. (Currently, server farms have been identified as one of the major electricity consumers in the world) [6]. Optimal resource allocation methods in datacenters can save up to 20% of the energy consumption. These savings may lead to an additional 30% saving in cooling energy requirements [6]. Given the rising cost of energy, cloud providers are presently looking for state-of-the-art solutions to ensure optimality of their power

consumption. Service of a job at a datacenter results in power consumption because of the server processing and network communication demands of this processing.

Hence, one option can be VM placement in a way that minimizes the number of active servers and communication traffic among VMs at the same time in favor of cloud computing DCs. Dynamic power management aims to reduce power consumption in DCs by temporarily shutting down servers when they are not required. However, full reactivation of a server is delayed by setup time, which can adversely affect the system performance. Hence, in order to be able to manage the number of active servers dynamically, the amount of incoming workload needs to be determined and assign VMs to new arriving jobs. This VM assignment process includes all server resources; namely CPU, memory and storage; which leads to a “multidimensional bin packing problem”.

Beside power usage of servers, communication also impacts both performance and cost of the operations. Communication increases job execution latency and power consumption. One way to mitigate the Cloud Network (CN) power usage is to apply traffic aware VM placement methods [7], [8]. Thus, the optimal solution of VM placement problem would include both CN and servers power consumption.

Literature Review

Cloud computing has emerged as a paradigm-shifting technology that provides on-demand access to a shared pool of computing resources. Effective resource allocation is crucial for achieving optimal performance and efficiency in cloud computing systems. Performance modeling and optimization techniques play a vital role in determining the appropriate allocation of resources, ensuring service quality, and maximizing resource utilization. This literature review provides an overview of key research contributions in the field of performance modeling and optimization of resource allocation in cloud computing systems.

Performance Modeling:

- Chaisiri et al. (2012) proposed a performance modeling approach based on the Stochastic Activity Networks (SANs) to predict the performance of cloud computing systems. Their study demonstrated that SANs can effectively capture the dynamic behavior of cloud resources and estimate system performance under various workload scenarios.
- Menasce et al. (2014) presented a performance modeling technique using queuing network models to evaluate the impact of resource allocation strategies on response time in cloud environments. They highlighted the importance of considering the variability in workload and resource availability when modeling cloud systems.
- Rimal et al. (2015) introduced a performance modeling framework that combines queuing theory and discrete event simulation to evaluate the performance of resource allocation policies in cloud environments. Their study emphasized the need for accurate modeling to identify performance bottlenecks and optimize resource allocation decisions.

Resource Allocation Optimization:

- Beloglazov et al. (2012) proposed an autonomic resource allocation technique based on the Ant Colony Optimization (ACO) algorithm. Their approach focused on optimizing the allocation of virtual machines (VMs) to physical servers in order to minimize energy consumption while meeting user-defined Quality of Service (QoS) constraints.
- Zhao et al. (2014) presented a multi-objective optimization approach for resource allocation in cloud systems. They formulated the problem as a multi-objective optimization problem and used the Non-dominated Sorting Genetic Algorithm II (NSGA-II) to find the Pareto-optimal solutions, considering objectives such as cost, energy consumption, and performance.
- Zhang et al. (2017) proposed a reinforcement learning-based approach for resource allocation optimization in cloud computing systems. They developed a Q-learning algorithm to dynamically allocate resources based on historical performance data and user demands. Their results demonstrated improved resource utilization and performance compared to traditional static allocation methods.

Performance-aware Scheduling:

- Wang et al. (2016) presented a performance-aware task scheduling algorithm for cloud computing systems. Their approach integrated both task scheduling and resource allocation decisions to minimize the makespan and improve resource utilization. The algorithm utilized particle swarm optimization to find the near-optimal solution in a reasonable time frame.
- Wei et al. (2018) proposed a performance-aware VM placement algorithm that considers both inter-VM interference and resource utilization. Their approach utilized an enhanced Differential Evolution (DE) algorithm to optimize VM placement decisions, taking into account the dynamic resource demands and heterogeneity of workloads.
- Yu et al. (2020) introduced a performance-aware task scheduling framework that considers both performance constraints and energy efficiency in cloud computing systems. Their approach employed a genetic algorithm to optimize task scheduling decisions and dynamically adjust resource allocation based on workload characteristics.

The literature review highlights the significance of performance modeling and optimization techniques in resource allocation for cloud computing systems. Various approaches such as stochastic modeling, queuing theory, optimization algorithms, and machine learning have been employed to

improve resource utilization, minimize response time, and enhance system performance. The reviewed studies emphasize the need for accurate modeling, dynamic resource allocation, and consideration of multiple objectives in optimizing resource allocation decisions.

Material and Methodology

Existing system: The Existing system is inspired by the concepts of the “Open Market” where any user can join, sell and buy their compute services. Cloud market conducted a double auction to match bids of multiple resource requests to the providers’ ask. First, the providers advertise their resources with their price (asks). Consumers submit their bids to show their interest in leasing the advertised resources. All the bids and asks are stored in the database which will be accessed at the end of the auction for calculating the winning bids. The Meta-Broker, which is the main agent of market, coordinates the matching of asks and bids, and trading between auction participants. At the end of the auction, the Meta-Broker decides the winners and sends the reservation requests to the Reservation Service of market. Then the Reservation Service informs the resource providers and consumers about the auction result. The information about reservations is stored within market using the Accounting service. [7]

Proposed system: The proposed system is an implementation of cloud market which shows that number of users is requesting for the resource, and cloud provider will provide the resource accordingly, in this prices for different resources are considered to be fixed but the best part of this system is if two users are requesting for the same resource with same price for the same time, the cloud provider will provide the resource to the one who is the loyal or trust worthy user. This can be calculated based upon the number of visits of user to the cloud market. In this we are considering the budget constraint i.e. if the user do not have appropriate budget then he is not allowed to request for a resource.

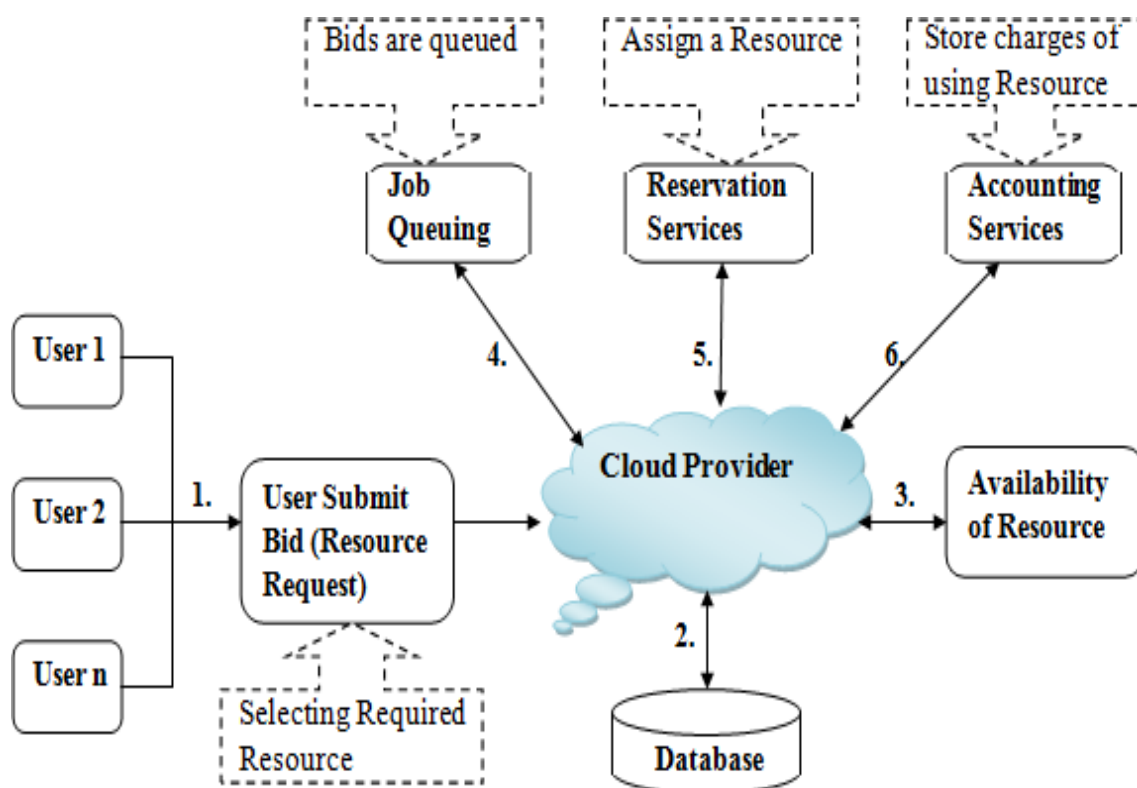


Figure 3.1 System Model for Resource allocation in cloud computing

User: Users are the consumers who submit a resource request to get a resource of their use. Consumers have to bid for a resource to show their interest in it.

- Cloud Provider: cloud provider is a central authority who accepts the requests of the users and checks the trading between auction participants. Cloud Provide the services to consumers after interacting with database.
- Database: Database is a back end of whole criteria; the detailed information like time, cost and budget of consumer is stored in it. Database allows the cloud provider whether to accept or reject the proposal of resource being used.
- Reservation Services: Users must register themselves before requesting for a resource. They are not allowed to access any of the resource before login. After registering they are assigned with the required resource.
- Job Queuing: Cloud provider maintains the no. of jobs by interacting with database; jobs are queued up according to bidding cost. The cloud provider will provide the resource to the one who is having the highest bid.
- Accounting Services: Accounting services maintains the information about the charges that a user has to pay for using a particular resource with respect to time

3.1 Sequence Diagram

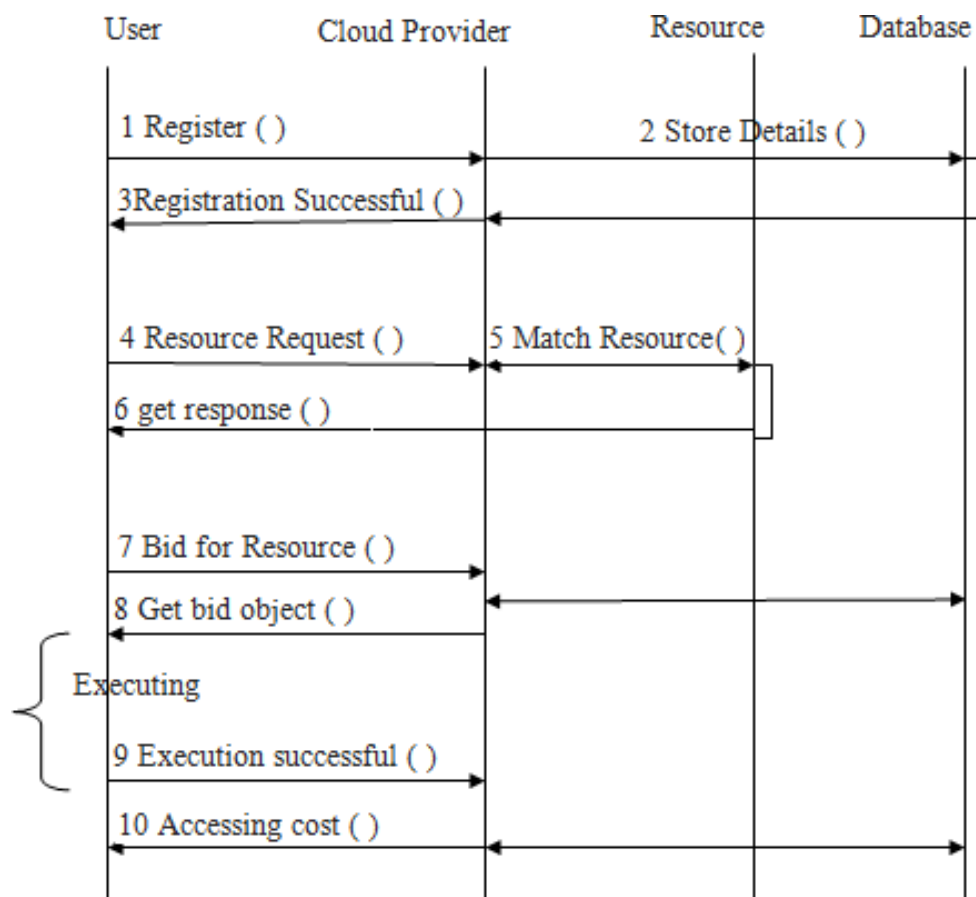


Figure 3.2 shows the sequence diagram for system model

Results and Discussions

In order to evaluate the performance of cloud market and provide a proof of concept of its architecture, we implemented a prototype and tested as a service provider. In this section, we describe the prototype implementation and discuss its performance results.

Market exchange has been implemented in Java in order to be portable over different platforms such as the Windows operating systems. From the implementation point of view, cloud market is composed of a collection of services that interact by means of a persistence layer represented by the SQL database. The system is accessible from external components through a web service that has been deployed by using Glassfish server. The web service interface makes the interaction with market as platform independent.

In this implementation, prices are considered to be fixed. It is more reliable since users can know more quickly whether he will get resources or not. To show the results, we use three metrics:

Time taken by cloud provider to assign the resource.

Cost incurred per consumer.

Time taken by cloud provider to assign the resource: when a user sends a resource request, cloud provider first checks whether the resource is available or not. Then it checks the bidding cost that a user submits for required resource and assigns the resource to the one having the highest bid

Table 4.1 shows the time taken by cloud server to assign a resource for execution

Client ID	Time Taken
1	7
2	7
3	6
4	8
5	10

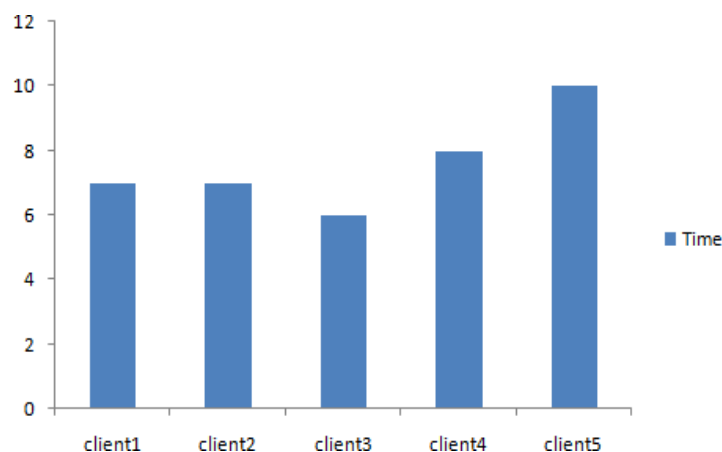


Figure 4.1 Performance of cloud Market for 5 clearance requests The graph shows the time between request and assigning the service to the client.

- a) Cost Incurred per consumer: As we know cloud provides the services “pay-as-you-go” basis. So consumers are charged with amount that they have to pay for using the resource

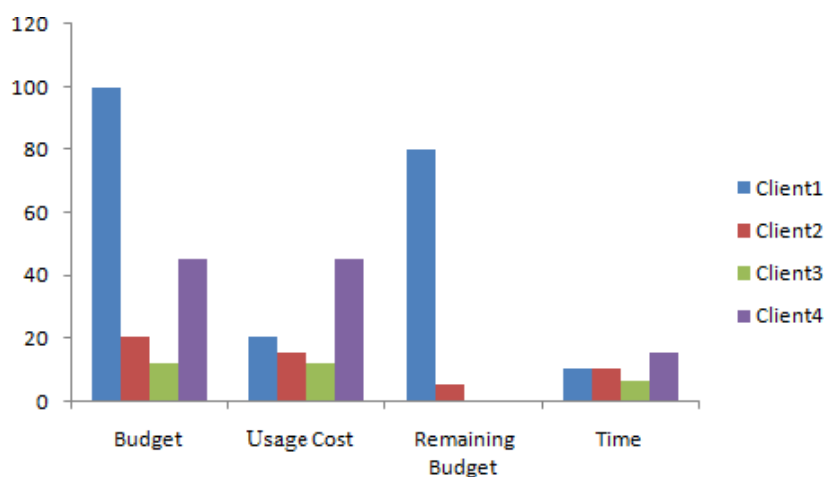


Figure 4.2 shows the graphical representation of user's detail

Cloud computing systems are popular that rent computing resources on-demand, bill on a pay-as-you-go basis, and multiplex many users on the same physical infrastructure. These cloud computing environments provide an illusion of infinite computing resources to cloud users so that they can increase or decrease their resource consumption rate according to the demands. Basically cost is according to the bulk of time that client had selected, if client wants more amount of time for single execution then obviously he has to pay more.

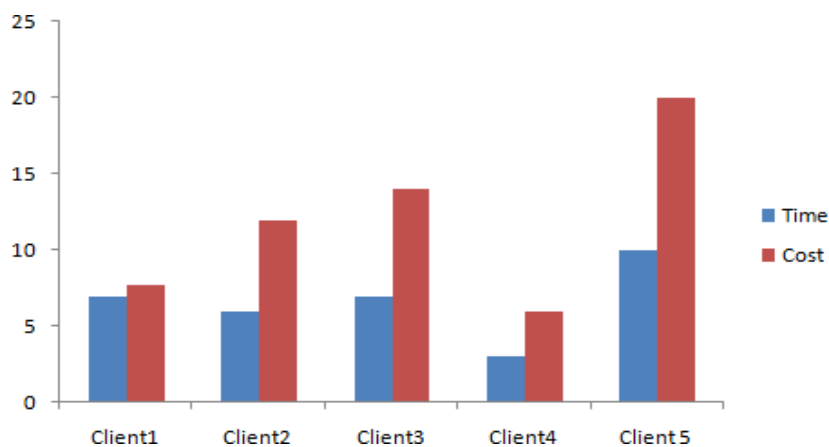


Figure 4.3 shows charges for executed services.

Table 4.2 shows time v/s cost graph

User	Time	Cost
Client1	7	7.7
Client2	6	12
Client3	7	14
Client4	3	6
Client5	10	20

Conclusion and Future directions

Conclusion

In this paper, we are considering the concept of budget constraint; cloud provider will provide the services based upon budget that the users use to get a resource. If user has appropriate budget then he will get the resource otherwise not. Budget is considered as a deadline and time keeps on changing according to the requirement of user. Moreover cost is considered to be fixed for a particular resource while time varies to execute a resource.

Future Scope

For future we will work on time constraint as well as by combining both time and budget constraint. In time constraint, time is considered to be deadline that user has to execute the resource in fixed time only otherwise cloud provider will de-allocate the resource forcefully. Secondly, both time and budget are considered as deadline, user has to execute his task within a particular amount of time with given budget.

REFERENCES

1. P. Mell and T. Grance. "The NIST definition of cloud computing", National Institute of Standards and Technology, vol. 53, no. 6, pp. 50, 2009.
2. S. S. Islam, M. B. Mollah, M. I. Huq, M. A. Ullah, "Cloud Computing for Future Generation of Computing Technology", Proceedings of the 2nd IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems, pp. 1-6, May 2012.
3. D. Jeffrey, and S. Ghemawat. "MapReduce: simplified data processing on large clusters." Communications of the ACM, vol. 51, no. 1, pp. 107-113, 2008.
4. T. White, "Hadoop: the definitive guide", O'Reilly, 2012.
5. C. Lam, "Hadoop in action", Manning Publications Co., 2010.
6. A. Berl, E. Gelenbe, M. Di Girolamo, G. Giuliani, H. De Meer, M. Quan Dang, and K. Pentikousis. "Energy-efficient cloud computing." The computer journal, vol. 53, no. 7, pp. 1045- 1051, 2010.
7. M. Alicherry, T.V. Lakshman "Network Aware Resource Allocation in Distributed Clouds", Proceedings of IEEE INFOCOM, pp. 963-971, 2012.
8. X. Meng "Improving the scalability of datacenter networks with traffic-aware Virtual Machine placement", Proceedings of INFOCOM, pp. 1-9, 2010.
9. Chaisiri, S., Lee, B. S., Niyato, D., & Wang, P. (2012). Optimization of resource provisioning cost in cloud computing. IEEE Transactions on Services Computing, 5(2), 164-177.
10. Menasce, D. A., Fonseca, R., Mendonca, N. C., & Sadok, D. (2014). Performance modeling and analysis of cloud services. IEEE Transactions on Services Computing, 7(3), 434-447.
11. Rimal, B. P., Choi, E., & Lumb, I. (2015). A taxonomy and survey of cloud computing systems. In 5th International Joint Conference on Cloud Computing (CLOUD) (pp. 44-51). IEEE.
12. Beloglazov, A., & Buyya, R. (2012). Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers. Concurrency and Computation: Practice and Experience, 24(13), 1397-1420.
13. Zhao, Q., Niu, L., Chen, H., & Chen, S. (2014). Multi-objective optimization for energy-efficient virtual machine placement in cloud data centers. Journal of Parallel and Distributed Computing, 74(10), 2891-2904.
14. Zhang, Q., Yan, Y., Wang, G., & Vasilakos, A. V. (2017). QoS-aware resource allocation for virtual machine-based cloud systems. IEEE Transactions on Parallel and Distributed Systems, 28(6), 1655-1669.
15. Wang, Y., Liu, X., Wu, L., Zhang, J., & Jin, H. (2016). Performance-aware task scheduling with multiple resource constraints for cloud computing. Future Generation Computer Systems, 56, 34-46.

-
16. Wei, J., Li, K., & Bao, W. (2018). Performance-aware VM placement in cloud computing: A differential evolution-based approach. *Future Generation Computer Systems*, 80, 29-39.
 17. Yu, L., Wang, J., & Tao, H. (2020). Performance-aware energy-efficient task scheduling in cloud computing. *Future Generation Computer Systems*, 102, 1084-1094.