International Journal of Interdisciplinary Innovative Research &Development (IJIIRD)
ISSN: 2456-236X
Vol. **05** Issue **02** | **2021**

A Survey on Resource Scheduling in Cloud Computing: Issues and Challenges

[1]Preethi.S, [2]Ganeshan M
[1]*MCA Scholar, School of CS & IT, Dept. of MCA, Jain (Deemed-to-be University)-069*
[2]*Assistant Professor, School of CS & IT, Dept. of MCA, Jain (Deemed-to-be University)-069*

ABSTRACT

Resource scheduling in cloud is a challenging job and the scheduling of appropriate resources to cloud workloads depends on the QoS requirements of cloud applications. In cloud environment, heterogeneity, uncertainty and dispersion of resources encounters problems of allocation of resources, which cannot be addressed with existing resource allocation policies. Researchers still face troubles to select the efficient and appropriate resource scheduling algorithm for a specific workload from the existing literature of resource scheduling algorithms. This research depicts a broad methodical literature analysis of resource management in the area of cloud in general and cloud resource scheduling in specific. In this survey, standard methodical literature analysis technique is used based on a complete collection of 110 research papers. The current status of resource scheduling in cloud computing is distributed into various categories. Methodical analysis of resource scheduling in cloud computing is presented, resource scheduling algorithms and management, its types and benefits with tools, resource scheduling aspects and resource distribution policies are described. The literature concerning to thirteen types of resource scheduling algorithms has also been stated. Further, eight types of resource distribution policies are described. Methodical analysis of this research work will help researchers to find the important characteristics of resource scheduling algorithms and also will help to select most suitable algorithm for scheduling a specific workload. Future research directions have also been suggested in this research work.

Keywords: ***Resource scheduling algorithms, Resource management, Resource distribution policies, Cloud computing, Resource scheduling tools, Cloud workloads, Resource scheduling aspects, Resource provisioning, Cloud resource scheduling***

1. INTRODUCTION AND MOTIVATION

Resource management is an umbrella activity comprising of different stages of resources and workloads from workload submission to workload execution. Resource management in Cloud includes two stages: resource provisioning and resource scheduling.

Resource provisioning is defined to be the stage to identify adequate resources for a given workload based on QoS requirements described by cloud consumers whereas resource scheduling is mapping an execution of cloud consumer workloads based on selected resources through resource provisioning as shown in Fig. 1. Firstly, cloud consumer submits request for workload execution in the form of workload details. Based on these details broker finds the suitable resources for a given workload and determines the feasibility of provisioning of resources based on QoS requirements. Broker sends requests to resource scheduler for scheduling after successful provisioning of resources. Other responsibilities of broker include: release of extra resources to resource pool, contains information of provisioned resources and monitor performance to add or remove resources. After resource provisioning, resource scheduling is done in second stage. All the provisioned resources are kept in resource queue while other remaining resources are in resource pool. Submitted workloads are processed in workload queue. In this stage, scheduling agent maps the provisioned resources to given workloads, execute the workloads and release the resources back to resources pool after successful completion of workloads. Based on QoS requirements, scheduling of resources for adequate workloads is a challenging issue. For an efficient scheduling of resources, it is necessary to consider the QoS requirements. There is a need to uncover the research challenges in resource scheduling to execute the workloads without affecting other QoS requirements. Resource scheduling is a hotspot area of research in cloud due to large execution time and resource cost. Different resource scheduling criteria and parameters are directed to different categories of Resource Scheduling Algorithms (RSAs). This research work discusses the second stage of resource management i.e., resource scheduling. Effective resource scheduling reduces execution cost, execution time, energy consumption and considering other QoS requirements like reliability, security, availability and scalability. In cloud environment, cloud consumer and cloud provider are two parties. Cloud consumer submits workloads while cloud provider provides resources for execution of workloads. Both the parties have different requirements: provider wants to earn as much profits as possible with lowest investment and maximize

utilization of resources while consumer wants to execute workloads with minimum cost and execution time. However, executing number of workloads on one resource will create interference among workloads which leads to poor performance and reduces customer satisfaction. To maintain the service quality, providers reject the requests that result in unpredictable environment . Providers also consider unpredictable resources for scheduling and execution of the workloads. Scheduling of resources becomes more challenging because both user and providers are not willing to share information with each other. The challenges of resource scheduling include dispersion, uncertainty and heterogeneity of resources that are not resolved with traditional RSAs in cloud environment

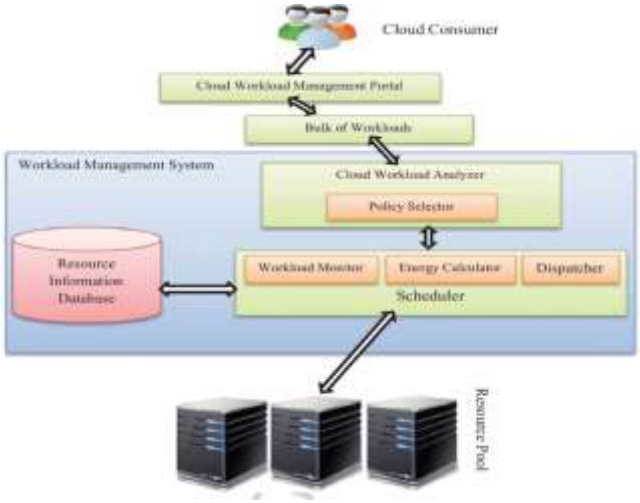

Fig 1: Resource scheduling in cloud

As shown in Fig.1, the workload details are gathered through the Cloud Workload Management Portal from cloud consumer. The number of cloud workloads submitted by the cloud user is processed in the queue. Based on the details given by cloud consumer, the resources are assigned to the cloud workloads for their execution. Resource provisioner provides the demanded resources to the workload for their execution in cloud environment only if required resources are available in resource pool. If the required resources are not available according to QoS requirement then the Workload Management System (WMS) asks to resubmit the workload with new QoS requirements in the form of SLA. After successful provisioning of resources, workloads are submitted to resource scheduler. Then the resource scheduler will ask to submit the workload for provisioned resources. After this resource scheduler will send back the results to WMS, cloud workload contains the resource information. Policy selector is used to select the appropriate scheduling policy based on workload details described by cloud consumer. Cloud environment and a scheduler that implements different scheduling policies based on the decision taken by policy selector. Based on the scheduling policy, the resources are allocated to the cloud workloads.

The resource scheduler schedules the incoming cloud workloads based on the workloads' details. First of all, get cloud workloads to schedule and then find appropriate and available resources and cloud workloads mapped efficiently based on the scheduling policies. Dispatcher is used to dispatch the workloads for execution. The workload is dispatched only, if the workloads will be executed according to the QoS parameters mentioned in SLA. Resource monitor is used to check the status of scheduling of resources like whether the required number of resources is provided or not. QoS monitor contains the information regarding QoS parameters to check whether all the workloads are executing within their specified range or not. Suppose deadline is a QoS parameter, so responsibility of QoS monitor is to check whether workloads are executed before desired deadline or not. There is violation of SLA if workload executes after desired deadline.

1.1 Need of Resource Scheduling

The first objective of resource scheduling is to identify the suitable resources for scheduling the appropriate workloads on time and to increase the effectiveness of resource utilization. In other words, the number of resources should be minimum for a workload to maintain a required level of service quality, or minimize workload completion time of a workload. For better resource scheduling, best resource workload mapping is required. The second objective of resource scheduling is to identify the adequate and suitable workload that supports the scheduling of multiple workloads, to be capable to fulfil numerous QoS requirements such as CPU utilization, availability, reliability, security etc. for cloud workload. Therefore, resource scheduling considers the execution time of every distinct workload, but most importantly, the overall performance is also based on type of workload i.e. with different QoS requirements (heterogeneous workloads) and with similar QoS requirements (homogenous workloads).

1.2 Motivation for Research

- Resource scheduling in cloud is a process of dynamic allocation of resources to cloud workloads after resource provisioning. Consequently, this study emphasizes on resource scheduling algorithms based on different scheduling criteria.
- We recognized the necessity of methodical literature survey after considering progressive research in resource scheduling in cloud computing. Therefore, we have summarized the available research based on broad and methodical search in existing database and present the research challenges for advanced research.

1.3 Related Surveys

The two researchers Vijindra et al. and Jose et al. have done innovative literature reviews in the field of resource scheduling. Still the research has persistently grown in the field of resource scheduling. There is a necessity of methodical literature survey to evaluate and integrate the existing research presented in this field. This research presents a methodical literature survey to evaluate and discover the research challenges based on available existing research in the field of resource scheduling in cloud computing.

2. BACKGROUND

In the beginning, we categorize the different types of resource scheduling algorithms and aspects leading to cloud resource scheduling.

2.1Resource Management

Cloud computing offers provisioning and scheduling of resources and provides guaranteed and reliable cloud services on the basis of pay per use policy. Due to fluctuation in demand of various cloud consumers, it is very difficult to provide the service in an effective way. We have identified the various resource scheduling techniques from the existing literature. The resource management in cloud is done for two stages: resource provisioning and resource scheduling as shown in Fig. 2.

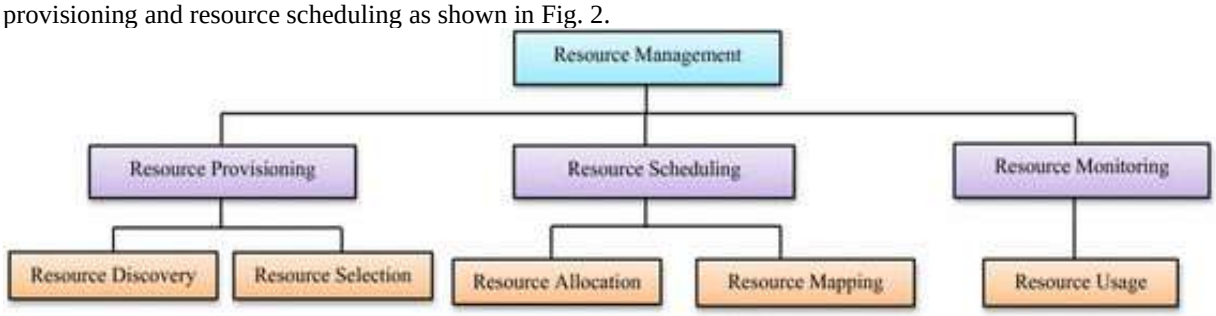

Fig.2 Taxonomy of resource management

Process of resource management is controlled by a centralized agent called Cloud Resource Manager (CRM). CRM manages all the cloud workloads and resources and maps the resources and workloads efficiently. There are different entities and interfaces associated with CRM as shown in Fig. Scaling listener is used to map the workloads with appropriate resources based on the QoS requirements as described by user. Generally in resource management, cloud consumer submits workloads along with their QoS requirements to the cloud provider for execution.

3. CLOUD RESOURCE SCHEDULING: OPEN ISSUES AND CHALLENGES

Though a lot of progress has been achieved and scalable computing infrastructures is easily implemented by cloud computing on pay per use basis. Still there are many issues and challenges in this field that need to be addressed. Based on existing research in cloud resource scheduling, we have identified various open issues still pending in this area. Research challenges based on these open issues have further been classified based on resource scheduling algorithms. Following open challenges and issues have been identified from the existing literature of resource scheduling in cloud computing:

3.1 Resource Scheduling

The challenges of resource scheduling include dispersion, uncertainty and heterogeneity of resources that are not resolved with traditional resource management mechanisms in cloud environment Thus, there is a need to make cloud services and cloud-oriented applications efficient by taking care of these properties of the cloud environment. Aim of resource scheduling is to allocate appropriate resources at the right time to the right workloads, so those applications can utilize the resources effectively. In other words, the amount of resources should be minimum for a workload to maintain a desirable level of service quality, or maximize throughput (or minimize workload completion time) of a workload. To address this problem, new solutions need to be developed.

3.2 Autonomic Resource Scheduling

Autonomic management implies the fact that the service is able to self-manage as per its environment. Autonomic management system is required for dynamic resource provisioning to fulfil the QoS requirements as described by cloud user and to reduce service cost and improve efficiency of the system. Cloud computing is an effective platform to execute web-based services on pay as-you-go basis but due to larger variation user demand, it is difficult to provision resources effectively.

3.3 Quality of Service (QoS)

To fulfil the QoS requirements of cloud service, required number of resources are provisioned by service provider. Based on these QoS requirements, SLA is designed and SLA violations are detected regularly, which further decides the penalty or compensation in case of SLA violation. Thus, there is need to provision adequate number of resources dynamically by service provider to reduce or avoid SLA violations.

3.4 Service Level Agreements (SLAs)

There is a need of autonomic cloud infrastructures to fulfil the QoS requirements described by cloud user in terms of SLA and to reduce interaction of cloud consumer with the computing environment. Therefore, effective strategy to detect SLA violation in advance is research issue which can avoid performance degradation.

3.5 Self-management Service

The aim of a cloud provider in this case is to assign and release resources from the cloud to fulfill its SLOs (Service Level Objectives), reducing its deployment charge. These methods usually include:

- Creating an application performance model that forecasts the number of application instances needed to manage request at every individual level, in order to fulfill QoS requirements;
- Occasionally forecasting forthcoming demand and defining resource requirements using the performance model; and
- Automatically assigning resources using the forecast resource requirements. The proactive method uses forecast demand to occasionally assign resources before resources are required. The reactive method responds to instant demand variations before periodic demand forecast is accessible.

3.6 Virtual Machine Migration and Server Consolidation

Virtualization can deliver important profits by allowing Virtual Machine (VM) migration to stable workload across the data center. Further, VM migration permits strong and highly responsive providing in data centres. Researchers have found that moving a whole OS and all of its applications as one unit allows avoiding many of the problems tackled by process-level migration methods, and investigated the advantages of migration of VMs. Detecting workload hotspots and initiating a migration lacks the agility to respond to sudden workload changes. Server consolidation is an operative method to improve resource utilization by reducing energy consumption. Energy can be saved through VM migration. To combine VMs, VMs should be located on many under-utilized servers onto a single server

3.7 Energy Management

In cloud computing, the improvement in energy efficiency is one of the major problems. It has been assessed that the price of powering and refrigeration accounts for 53 % of the entire operational spending of data centres. In 2006, data centers in the US consumed more than 1.5 % of the total energy produced in that year, and the proportion is estimated to grow 18 % yearly. Therefore, infrastructure providers are under huge pressure to decrease energy consumption. The aim is not only to decrease energy cost in data canters, but also to meet government rules and environmental standards. Energy-oriented task scheduling and server consolidation are two other methods to decrease power consumption by switching off free systems. A main issue in existing techniques is to attain a decent trade-off between energy savings and application performance.

3.8 Data Security

Data security is another open issue in cloud computing. Meanwhile cloud providers usually do not have access to the physical data security system of data centers cloud provider must depend on the infrastructure provider to attain complete data security. Even for a virtual private cloud, the cloud provider can only identify the security setting distantly, without knowing whether it is completely implemented or not. It is dangerous to form trust procedures at each architectural layer of the cloud. Initially, the hardware layer must be reliable using hardware reliable platform segment. Furthermore, the virtualization platform need be confidential using secure VM observers. VM migration should only be permitted if both sender and receiver servers are confidential.

3.9 Dynamic Scalability

The aim of scaling and resource scheduling is to maximize application performance within budget constraints in cloud workloads. What resources should be acquired/released in the cloud, and how should the computing activities be mapped to the cloud resources, so that the application performance can be maximized within the budget constrains? Dynamic scalability is the ability to acquire or release the resources in response to demand dynamically. In a data center, the primary goal of a dynamic autonomous resource management process is to avoid wasting resources as a result of under-utilization.

3.10 Benefits of Cloud Resource Scheduling

We found numbers of benefits of cloud resource scheduling from existing literature, some important of them are:

- Effective cloud resource scheduling increases the robustness and minimizes makespan of workflow simultaneously.
- Reduce execution time and computation time of cloud workloads in effective cloud resource scheduling.
- Better resource utilization under different requirements of priority and avoid under loading and over loading of resources.
- No scheduling delay and lesser chances of resource failure due to efficient allocation of resources.
- No long VM startup delay, schedule provisioned resources immediately in effective cloud resource scheduling.
- Meet even strict application deadline with minimum budget expenditure and increases global profit in effective cloud resource scheduling
- Power consumption reduced without violation of SLA in effective cloud resource scheduling.
- Improve user deadline violation rate due to resource scheduling after resources provisioning.
- Waiting time is lesser of workloads on queue in effective cloud resource scheduling.
- Minimize carbon footprints and enabled dynamic scalability to handle demand fluctuation in effective cloud resource scheduling.
- Provide robust node for heterogeneous services, less chances of unplanned failure, no negative impact on server performance and node resource utility.
- Efficient balancing load by efficiently distributes the workloads on available resources in effective cloud resource scheduling.

4. CONCLUSIONS

In this research paper, the results have been analyzed in various ways like classification of resources, resource scheduling evolution as per research questions, percentage of different scheduling algorithms and their QoS parameters, detailed classification of resource scheduling algorithms and their subtypes, comparison of resource scheduling algorithms, resource scheduling aspects, resource distribution policies and resource scheduling tools have been presented. Recent research has shown that resource scheduling algorithms using resource provisioning mechanisms. Also it is not easy to find the best mapping of workload and resource without effective resource provisioning technique. Based on existing research, there is need of proper understanding of QoS requirements of workload for better resource allocation instead of detecting workload and resource. There is a need to finding the progresses in cloud research itself before find the advance research in resource scheduling.

5. REFERENCES

[1] Kc, K., Anyanwu, K.: Scheduling hadoop jobs to meet deadlines. In: 2010 IEEE Second International Conference on Cloud Computing Technology and Science (CloudCom), pp. 388–392. IEEE (2010)

[2] Lee, Y.C., Wang, C., Zomaya, A.Y., Zhou, B.B.: Profitdriven service request scheduling in clouds. In: Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, pp. 15–24. IEEE Computer Society (2010)

[3] Hu, J., Gu, J., Sun, G., Zhao, T.: A scheduling strategy on load balancing of virtual machine resources in cloud computing environment. In: 2010 Third International Symposium on Parallel Architectures, Algorithms and Programming (PAAP), pp. 89–96. IEEE (2010)

[4] Pandey, S., Wu, L., Guru, S.M., Buyya, R.: A particle swarm optimization-based heuristic for scheduling workflow applications in cloud computing environments. In: 2010 24th IEEE International Conference on Advanced Information Networking and Applications (AINA), pp. 400–407. IEEE (2010)

[5] Yang, Z., Yin, C., Liu, Y.: A cost-based resource scheduling paradigm in cloud computing. In: 2011 12th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT), pp. 417–422. IEEE (2011)
[6] Wu, L., Garg, S.K., Buyya, R.: SLA-based resource allocation for software as a service provider (SaaS) in cloud computing environments. In: 2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), pp. 195–204. IEEE (2011)
[7] Ying, C., Jiong, Y.: Energy-aware genetic algorithms for task scheduling in cloud computing. In: 2012 Seventh ChinaGrid Annual Conference (ChinaGrid), pp. 43–48. IEEE (2012)
[8] Sotomayor, B., Montero, R.S., Llorente, I.M., Foster, I.: Virtual infrastructure management in private and hybrid clouds. IEEE Internet Comput. 13(5), 14–22 (2009)
[9] Van den Bossche, R., Vanmechelen, K., Broeckhove, J.: Cost-optimal scheduling in hybrid iaas clouds for deadline constrained workloads. In: 2010 IEEE 3rd International Conference on Cloud Computing (CLOUD), pp. 228–235. IEEE (2010)
[10] Oprescu, A., Kielmann, T.: Bag-of-tasks scheduling under budget constraints. In: 2010 IEEE Second International Conference on Cloud Computing Technology and Science (CloudCom), pp. 351–359. IEEE (2010)